SPATIAL POINT PROCESS MODELLING OF A BIODIVERSE PLANT COMMUNITY

JANINE BÄRBEL ILLIAN

A thesis submitted in partial fulfilment
of the requirements of the
University of Abertay Dundee
for the degree of Doctor of Philosophy

February 2006

I certify that this thesis is the true and accurate version of the thesis approved by the examiners.

Signed... [redacted] ..... Date.....21/6/06.....

(Director of Studies)

Abstract

The fundamental question of this thesis addresses the suitability of spatial point processes for the modelling of plant communities with the aim of understanding the mechanisms that sustain biodiversity and allow species coexistence through an analysis of a biodiverse plant community in Western Australia.

Terrestrial plants are non-motile organisms and hence interact mainly with their nearest neighbours. Therefore, plant community dynamics need to be modelled in a spatial context. Spatial point process models describe the arrangement of objects in space and derive information on inter- and intra-species interaction through an analysis of the objects' relative position. However, there has been a lack in appropriate methodology since previous applications of spatial point processes have analysed data sets of much smaller complexity and research has been mainly theory driven.

This thesis develops new methodology suitable for this context. It provides parsimonious tools for the exploratory data analysis and derives multivariate methods for spatial point pattern data for the first time. In addition, three models of increasing complexity are fitted to the data and their suitability is assessed.

This thesis has provided a number of statistical tools that may also be successfully applied in other situations, notably where highly multivariate data sets of spatial patterns occur. The methods are suitable in the context of plant community dynamics and their application to the study data set has made contributions to the development and validation of existing eco-

logical theories on biodiversity. Furthermore, they informed specifically on intra- and inter-species interaction structures in the study data set and are ultimately contributing to conservation.

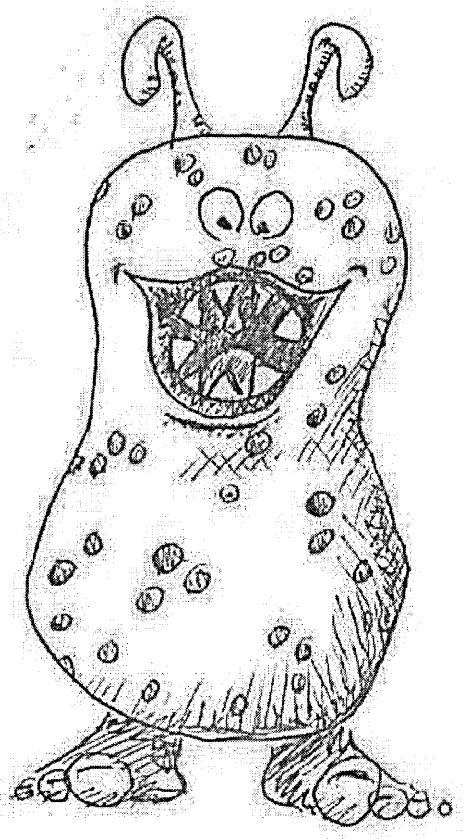

Figure 1: A Punktmönster...

Figure 2: Flowers of *Banksia grandiis* and cone of *Banksia grandiis*; second photograph courtesy of Paul Armstrong

Insofern sich die Sätze der
Mathematik auf die Wirklichkeit
beziehen, sind sie nicht sicher,
und insofern sie sicher sind,
beziehen sie sich nicht auf die Wirklichkeit.

(Albert Einstein)

Hat die Natur sich auch verschlechtert,
Und nimmt sie Menschenfehler an?
Mich dünkt, die Pflanzen und die Tiere,
Sie lügen jetzt wie jedermann.
(Heinrich Heine, *Entartung*, aus: *Zeitgedichte*)

La Nature est un temple où de vivants piliers
Laissent parfois sortir de confuses paroles;
L'hommes y passe à travers des forêts de symboles
Qui l'observent avec des regards familiers.

(Charles Baudelaire, *Les Fleurs du Mal*)

Acknowledgements

I owe an immense debt of gratitude to many people without whom I would have never been able to finish this thesis. I would like to thank each and everyone of them for their incredible patience.

First and foremost, I am greatly indebted to my director of studies Harry Staines for all his encouragement, support and understanding and, of course, for introducing me to the data and for the many things I have learned from him. I would also like to thank my supervisory team; Erica Benson for her personal support and for introducing me to the ecological and conservation aspects of the project, John Crawford for many thought-provoking discussions and for his generosity and Jesper Møller for welcoming me again and again to Aalborg and for his humour.

I would also like to express my special gratitude to Jim Bown for the countless things he did in the last years but mainly for the many challenging discussions which clearly had a strong impact on the contents and focus of this thesis.

Undoubtedly, this thesis would have never been possible without Paul Armstrong who not only collected the data but also made every effort to make my visits to Australia enjoyable and taught me so much about the

Australian flora. I would also like to thank Kingsley Dixon for providing his expert insight into the Australian flora, Rasmus P. Waagepetersen for many discussions and his invaluable assistance with the computational side of this thesis and Martin Schlather for his initial advice, discussions and for always pushing me in the right direction.

The many discussions I had with Richard Law, David Burslem and Manuel Schneider as well as the other members of the working group have opened my eyes to ecological modelling, plant community dynamics and questions of biodiversity that I would never have become aware of without them – thanks so much. I big thanks also goes to Dietrich Stoyan, Antti Pentinnen, Adrian Baddeley, Robin Milne and Paul Blackwell for scientific inspiration.

I am also indebted to Jan and Louis Natanson for the time we could spend in their wonderful house in Whitehills where a fair bit of this thesis was written in the summer of 2005.

Over the years of my studies I received a lot of support from staff in the School of Computing at Abertay University. A special thanks goes to Anne Farrow and Nat Jack as well as to all the other staff in the former Division of Mathematical Sciences for their constant understanding, and to my PhD student Jacqueline Birrell for her patience.

I would also like to thank Merete Heide for her very flexible and often last-minute help during my stays in Aalborg and for her friendship. A big hug to Jayanthi Muthusamy who has been a great companion and friend through all the years. Mange tak to Kasper K. Berthelsen for those never-ending conversations on basically everything and, of course, the punktmönster. Vielen lieben Dank to Volker Schmid for putting me up in London and for everything

else.

The research community within SIMBIOS has been extremely supportive. There are too many people to mention individually, but some names stand out; I would like to thank Ruth Falconer for all the runs and organising my farewell party, Alistair Eberst for his patience and the banter, Alasdair Houston for always having an answer to all my notorious questions.

Looking back, I would like to mention Jörg Voegele and Niels Waller who introduced me to research and ignited my scientific curiosity.

To all my other friends – whether they are in this country or in Germany or elsewhere in the world – I would like to say sorry for making myself so rare, especially in the last year.

Und ganz besonders möchte ich natürlich meinen Eltern Bärbel und Hermann Illian danken, ohne die ich niemals soweit gekommen wäre. Ein ganz besonders herzlicher Dank gilt auch meinem Mann Claus-Dieter Mayer, der mich von der ersten Diskussion in Crail bis zur Abgabe der Arbeit mit sehr viel Geduld begleitet hat.

In memory of Anne

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Ultimately, the earth as a system is dependent on the functioning of natural and managed ecosystems since it is regulated by the biogeochemical processes derived from them (Loreau et al. 2001). Recent decades have seen an increasing decline in species' biodiversity as a result of human interference (Cardinale et al. 2004; Regan et al. 2001). The potential ecological consequences of biodiversity loss have led to a growing concern about the future survival of ecosystems and their functioning (Mouquet et al. 2002). Consequently the relationship between biodiversity and ecosystem functioning constitutes a major scientific issue today (Loreau 2000; Tilman et al. 1996, 2001; Hector 1999). However, understanding the impact of biodiversity loss on function requires an understanding of the processes that organise ecosystem communities and the mechanisms that sustain biodiversity (Bell 2001; Hubbell 2001; Chave 2004; Condit et al. 2002; Duivenvoorden et al. 2002). Substantial research focuses on modelling structures and processes in plant ecosystems and communities. Plants are primary producers and hence represent the basal component of most ecosystems with many terres-

trial ecosystems' survival and diversity depending on the resources plants provide as well as on their structures and diversity (Loreau et al. 2001).

Key research in community ecology thus seeks to reveal the mechanisms that allow a large number of species to coexist (Murrell et al. 2001; Loreau et al. 2001). Coexistence primarily concerns the inter- and intra-specific interactions in a community (Durrett and Levin 1998). Since individual plants interact mainly with their neighbours (Tilman 1994; Dieckmann et al. 2000; Purves and Law 2003; Stoll and Weiner 2000) interactions between plants in plant communities typically take place in a spatial context and hence current modelling approaches consider individuals in spatially explicit models (Chesson 2000; DeAngelis and Gross 1992; Huston et al. 1988; Judson 1994).

Spatial point process models are statistical models that describe the exact locations of objects in space. They model the pattern formed by these objects based on the interactions among them and on their properties (such as their height), taking potential spatial heterogeneity into account (Diggle 1983; Cressie 1991; Møller and Waagepetersen 2003b). The spatial pattern formed by the individuals in a plant community may be a result of species interaction and environmental heterogeneity (Law et al. 2000; Armsworth et al. 2004). Spatial point processes may therefore be used as models of plant communities to infer interaction structures and dependence on local growing conditions.

This thesis uses spatial point processes to model the spatial pattern formed by individuals in a natural plant community with a high degree of biodiversity in Cataby, Western Australia (Armstrong 1991). The environmental conditions are unknown within the plot and may be considered homo-

geneous (Dixon 2005), and so this thesis focuses on modelling the locations of individuals within the community to analyse its interaction structure.

Most applications of spatial point processes so far have been theory driven rather than data driven (Møller and Waagepetersen 2003b). Hitherto they have not been applied to entire plant communities and at most two or three species have been modelled simultaneously (Diggle 2003). However, data sets of plant communities with high biodiversity, tend by definition to be much more complex. With 6378 plants from 67 species on a 22m $\times$ 22m plot, the data set considered here consists of the locations of a larger number of individuals from a much larger number of species than has been analysed so far. As a result, all aspects of the modelling process, such as parameter estimation, presentation and interpretation of results, present new challenges.

This thesis seeks to establish the suitability of the spatial point process methodology in the context of highly biodiverse plant community ecology. It develops new approaches that may be applied to inherently complex data sets with the aim of drawing conclusions from the results on species coexistence and biodiversity for the specific data set in particular and in general. It focuses on the following aspects:

a) Facilitating exploratory data analysis in a multi-species setting to analyse intra-species interactions;

b) Developing multivariate statistical methods with the goal of reducing complexity to jointly characterise the spatial behaviour in the community;

c) Constructing a parsimonious model that may be fitted to a complex data set to yield information on inter-species interaction.

This thesis is organised as follows. Chapter 2 establishes why spatial point process modelling may be considered a suitable approach in the context of plant communities. It details the ecological background necessary for interpreting the results of a spatial point process analysis with regard to community dynamics and biodiversity. Chapter 3 provides the statistical background through an overview of existing spatial point process methodology. In Chapter 4, novel methods are explored that are suitable for the initial descriptive analysis of a multi-species spatial pattern (see a) above). Chapter 5 presents multivariate statistical methodology, in particular principal component analysis for spatial point processes (see b) above). Chapter 6 develops three increasingly complex and realistic spatial point process models (see c) above). Chapter 7 critically discusses the methodology developed and the results obtained in Chapters 4 to 6 and details scope for further work. We will eventually be able to assess the suitability of spatial point process modelling for data sets of plant communities and their capability of informing biodiversity theory.

Chapter 2

Modelling plant biodiversity in a spatial context

Natural and managed habitats are undergoing unprecedented change, as a consequence of human activity, resulting in an unprecedented loss in biodiversity. To identify conservation strategies to counter this loss, ecologists therefore seek to understand the processes that organise ecosystem communities. This may help predict how disturbance and external changes to an ecosystem affect its dynamics (Armsworth et al. 2004). Furthermore, such knowledge may help the renaturation and subsequent conservation of disturbed ecosystems (Greig-Smith 1983; Herben et al. 2000). A fundamental question is whether depauperate systems function differently from or less efficiently than ecosystems and communities with higher biodiversity and, in the long run, whether loss of species with specific traits and functions will have an influence on productivity (Cardinale et al. 2000; Loreau et al. 2001).

Research has shown that biodiversity has positive short-term consequences for an ecosystem, leading to increased plant biomass production and nutrient retention (Tilman et al. 1996, 2001; Hector 1999). While diversity may not

be critical for maintaining ecosystem processes under benign environmental conditions, it may become much more important under changing more malign environmental conditions, and thus also have positive long-term effects for an ecosystem. Consequently, a given species may appear redundant at a given time, i.e. its function can be taken on by another species. It may not be redundant at some point in the future, however, when environmental conditions have changed and are maybe less benign for the overall community but probably more benign for the specific species (Loreau 2000). As a result, ecosystems with high levels of biodiversity may thus be more resilient to perturbation. The large species pool enables the system to maintain its productivity despite the loss of selected species that are unable to survive under changed growing conditions (Tilman and Downing 1994; Naeem and Shibin 1997), e.g. resulting from global warming (Petchey et al. 1999).

Consequently, revealing the mechanisms that allow a large number of species to coexist is of key interest within community ecology (Murrell et al. 2001; Loreau et al. 2001). Species coexistence, in turn, is directly linked to local inter- and intra-specific interactions in a community (Durrett and Levin 1998). These interactions typically take place in a spatial context, i.e. the environmental conditions at the spatial locations of the individuals and the local interaction structure among individuals in space have to be taken into account (Chesson 2000). This is particularly important in the context of ecological communities of sessile species such as terrestrial plants, as these interact mainly with their immediate neighbours due to their limited mobility (Tilman 1994; Dieckmann et al. 2000; Stoll and Weiner 2000; Purves and Law 2003).

This chapter establishes the relevance of considering space in ecological modelling and describes current non-statistical approaches to modelling species coexistence (Section 2.1). In Section 2.2, we argue why statistical methods may be of use in plant community ecology and review available spatial statistical methods. We establish why spatial point process modelling is the most suitable approach for the issues considered in this thesis. In Section 2.3, we consider spatial point process modelling of plant communities and how these methods may help the further development of theories of biodiversity. We also point out some of the limitations of current spatial point process approaches and make a case for the development of new methods necessary in the given context. In Section 2.4, we introduce a data set relating to a highly diverse community, collected in Southwestern Australia. We outline why the analysis of the data set may contribute to an understanding of biodiversity and argue that through using a statistical approach the analysis itself may help establishing the appropriateness of the novel methodology. We close by detailing the specific aims and objectives of this thesis (see Section 2.5).

2.1 Integrating spatial structure into ecological modelling

"Space is the final frontier" in ecology (Liebhold et al. 1993)

Whereas temporal aspects of ecological processes and the development of ecological communities over time have been studied extensively in the past since the works of Lotka and Volterra (Lotka 1925; Volterra 1926), spatial aspects were ignored for much longer (Perry et al. 2002). This may be partly

due to the resulting increased complexity and sheer magnitude of data (Liebhold and Gurevitch 2002), together with a lack of available methodology coupled with low computing power (Law et al. 2000). Technical advances have improved the performance of both computer hardware and software, and the development of geographical information systems (GIS) makes data more readily available. This has made a whole host of new, more complex modelling possibilities feasible.

In fact, the lack of spatial independence in ecological data was initially viewed as a problem that needed to be eliminated (Legendre 1993). Only in the last 20 years, however, has the spatial relationship itself been investigated by ecologists (Liebhold and Gurevitch 2002), with the aim of explicitly measuring and modelling the spatial patterns formed by organisms. The resulting general increase in the number of studies taking spatial aspects into account explicitly can be observed in the literature. We find this reflected in the growing interest in spatial scale (Wiens 1989; Dungan et al. 2002; Levin 1992), meta-population dynamics (Hanski and Gilpin 1997), spatio-temporal dynamics (Hassell et al. 1991), spatially explicit modelling (Silvertown et al. 1992) and spatial synchrony (Bjørnstad et al. 1999; Perry et al. 2002). In addition, it has been acknowledged that spatial modelling may promote the construction of spatial ecology theory and spatially explicit models (Hassell et al. 1991; Dunning et al. 1995; Tilman and Kareira 1997; Liebhold and Gurevitch 2002).

"The world is a patchy place " Dale (1999).

Environmental conditions are heterogeneous across all spatial scales (Levin

1992) such that most systems in the natural world do not show any spatial homogeneity but exhibit spatial structure (Dale et al. 2002). In plant communities, the observed spatial structure may be the result of both heterogeneous resource availability and inter- and intra-species interactions (Armsworth et al. 2004). Thus, analysing the spatial structure while taking spatial heterogeneity into account may aid in the understanding of inter- and intra-specific interaction, i.e. competition or facilitation (Perry et al. 2002). Due to their limited mobility, individual organisms interact mostly with their neighbours. This is particularly true for terrestrial plants and other sessile organisms (Tilman 1994; Purves and Law 2002; Stoll and Weiner 2000). Furthermore, individuals are often distributed in clumps and thus experience contact with close neighbours only (Durrett and Levin 1994; Purves and Law 2003). Individuals interact with neighbouring individuals either by inhibiting (negative interaction) or by facilitating (positive interaction) the growth of another individual in their vicinity. Negative interaction between neighbours may occur in the form of shading above ground and competition for nutrients and water below ground (Law et al. 2001), allelopathy (Crawley 1997b) and positive interaction in the form of pest facilitation or sub-canopy soil enrichment (Wright 2002; Bertness and Callaway 1994; Callaway 1995). An inspection of the spatial structure may also reveal complex local interactions between the species and the environment they live in (Dale 1999) and their influence on growth, and birth and death rates of individuals (Schneider et al. 2006).

Theoretical modelling has shown that spatial heterogeneity influences population and community dynamics, as models ignoring spatial structure produce substantially different results from those which do include spatial

dynamics (Tilman and Kareira 1997). Spatial heterogeneity has a major role in the persistence of species, the stability of populations and coexistence of species (Chesson 2000; Tilman 1994; Hanski and Gilpin 1997; Tilman et al. 1997; Bascompte and Sole 1998). Population persistence in heterogeneous environments may be considered the result of interaction between local density-dependence, dispersal and spatial heterogeneity (Amarasekare 2003). Thus, spatial structure has been invoked to explain competitive coexistence. When modelling the spatial structure of the pattern formed by a plant community we must define what we understand by a spatial pattern to avoid misunderstandings, since the term usage is inconsistent in the literature. E.g. Dale (1999) uses the following definition:

> " (...) When the patchiness has a certain amount of predictability so that it can be described quantitatively, we call it a pattern. Although the concept of pattern is often associated with non-randomness, in some cases we will want to allow the possibility of random pattern, because true randomness does permit a certain amount of prediction."

We do not completely agree with this definition, as we find it impossible to decide whether an observed spatial structure can be considered admissible to quantitative description. Furthermore, the inability to quantify a structure using a given method does not automatically imply that it cannot be quantified at all. Perhaps, an inappropriate model has been chosen, or suitable models have not yet been developed. Also, as we are intending to apply

statistical models we regard random patterns[1] as no less predictable than non-random patterns.

We define a *spatial pattern* to be the spatial formation of constituents of any type. In the context of plant ecology these constituents may be individuals, assemblies of individuals, communities or populations of plants. As a result of the ubiquitous heterogeneity mentioned above, patterns we find in nature tend to be non-random (Perry et al. 2002). For instance, plants in mesic environments are commonly clustered together whereas the spatial pattern of plants in arid environments may sometimes be regularly spaced, indicating repulsion among individuals (Purves and Law 2003; Silvertown et al. 1992).

2.1.1 Pattern and process

The main justification for analysing spatial patterns found in nature, and more specifically in plant communities, is that spatial characteristics and other processes such as establishment, growth, competition, reproduction and mortality of an individual are not independent from neighbouring individuals. As a consequence, we have two major aims when taking spatial structure into account:

(a) either modelling non-spatial characteristics of the individuals based on the underlying spatial dependence among individuals; or

[1]In Chapter 3, we will formally define what we mean by a "random" pattern. For now, we regard a pattern as spatially random when the locations and properties of all its constituents are independent of the locations and properties of all other constituents.

(b) modelling the spatial characteristics of plant communities taking potential environmental inhomogeneity into account, based on non-spatial characteristics and knowledge on underlying ecological processes (Dale 1999).

Section 2.2 notes that there are statistical methods available for analysing data in both situations, although in this thesis, we are primarily interested in the second aspect (i.e. in (b)).

However, caution must be exercised when linking pattern and process. In general, even though each pattern is a result of a process, it is not possible to link uniquely patterns and processes, since different ecological processes may lead to the same spatial pattern (Cale et al. 1989; Liebhold and Gurevitch 2002). Furthermore, structured or deterministic processes may also produce an apparently random pattern (Rummel and Roughgarden 1983; Case and Sidell 1983). Cale et al. (1989) go as far as suggesting that it is not the patterns that should be analysed but the fundamental processes themselves.

Nevertheless, we believe that the quantification of a spatial pattern may provide an indication of properties of underlying processes such as intra- and inter-specific interactions and interactions between individuals and the environment. The quantitative description yields a clear and objective description of spatial patterns and may be used to confirm or rule out specific hypothesised ecological processes (Dale 1999). For instance, the observation of increased inter-plant distances with time strongly contradicts a hypothesis of predominantly positive interactions in a community (Lepš 1990), whereas the aggregation of a specific species in areas with distinct environmental con-

ditions indicates microhabitat association (Harms et al. 2001; Burslem and Law 2005). However, it does not suffice to model merely the underlying processes on their own. One has to relate these to observed patterns in order to draw valid conclusions on natural processes. Thus, the approach taken in this thesis models an observed spatial pattern by using information on the potential underlying processes and assessing the strength of their influence on the pattern. This enables us to establish whether it is likely that assumed processes are actually ecologically relevant for the community, in as far as that they have contributed to the formation of the specific observed pattern.

2.1.2 The issue of scale

Most ecological processes are scale dependent and spatial characteristics vary across scales (Wiegand and Moloney 2004). These scales may range from a few millimetres to many kilometres, i.e. from the neighbourhood of individual plants to the scale of entire landscapes (Dale 1999). Depending on the scale chosen, different aspects of the underlying ecological processes that are related to the observed pattern are considered. More specifically, when modelling patterns at the smallest scales the interactions among individuals are reflected in the pattern and hence are related to population dynamics. At larger scales, on the other hand, the patterns may reveal interaction between animals and plants, e.g. through pollination and seed dispersal (Crawley 1997b).

In addition, when analysing a specific spatial pattern we need to bear in mind that we only perceive and model the pattern at the scale of data collection. Consider Figure 2.1, which shows an artificial example of the

pattern formed by the location of plants. It is clear that the overall pattern is clustered but that the pattern shows some form of regularity or repulsion at a smaller scale, e.g. in the smaller rectangle. If the data had only been collected within the small rectangle any information on the larger scale clustering would have been lost. Sometimes, additional information is available on larger scale properties, such that this can be taken into account. This is the case for the data set analysed in this thesis, where the observed plant community is known to be located in an area with large-scale patchiness. The data have been collected within one of these patches (see Section 2.4).

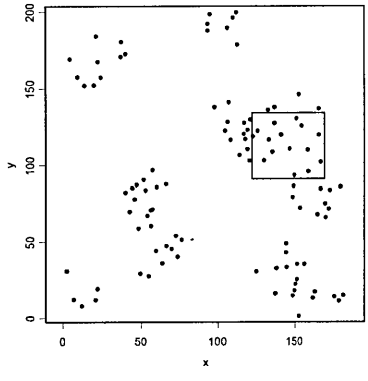

Figure 2.1: Example of a spatial pattern clustered at a large scale but regular at the scale of observation.

In summary, when describing or analysing a specific pattern we must recognize that we may only draw conclusions at the scale at which we work, and it is important that one is aware of this when analysing and modelling spatial patterns (Dungan et al. 2002).

2.1.3 Causes of patterns

As established above, different mechanisms operate at different scales often resulting in different types of patterns at different scales. This already indicates that an observed pattern formation may not be the result of a single factor but has been caused by a number of different factors operating simultaneously. We adapt an approach by Dale (1999) here, and subdivide the factors that affect and cause spatial patterns broadly into three groups: morphological factors, abiotic environmental factors and biotic or phytosociological factors. All these factors may be operating in any plant community but their relative relevance in general and for a specific community may vary. More specifically, through the analysis of its spatial pattern we may draw conclusions on these different types of factors and their relative importance for a given plant community.

Morphological factors These include the size and the growth pattern of the involved plants that influence an observed spatial pattern. For instance, Mahdi and Law (1998) point out that clonal growth has caused the spatial organisation of plants in a limestone grassland community. Similar conclusions have been drawn in Kershaw (1964) and Dale and MacIsaac (1989). Likewise, a pattern may have been influenced by the size of the involved individuals. Larger plants may repulse smaller plants through shading and nutrient uptake causing more regular patterns (Crawley 1997b).

Abiotic environmental factors There is a large number of abiotic environmental factors that have an impact on the spatial structure of plant

communities. These include properties of the soil, such as soil depth, soil nutrient level, position of surface rocks etc., as well as the water level and availability of light. Patterns may also be caused by topographic heterogeneity due to glaciation effects, ancient dunes, reef ridges, etc. (Dale 1999).

Another important abiotic environmental factor that leads to spatial patterning at both a very local scale but also a more global scale is the influence of abiotic disturbance, such as storms or fire. The latter is particularly relevant for the plant community analysed here, as the area is susceptible to regular fires, a fact that has directly influenced speciation and resulted in different regeneration mechanisms, see Section 2.4 for details. The modelling approach in Chapter 6 takes this into account explicitly.

Biotic environmental factors A main biotic environmental factor relevant for spatial patterning is the prevalent interaction among individuals, in the form of either intra-specific or inter-specific competition or facilitation (Law et al. 2001). Another major group of biotic environmental factors influencing spatial patterns are biotic disturbances, such as falling trees, etc. and the influence of foraging animals.

In the data set analysed in this thesis, soil nutrient and water levels are reported to be extremely low but nevertheless homogeneous within the selected site such that we can assume that most of the spatial structure has been caused by intra- and inter-specific competition for resources (Dixon 2005; Armstrong 2005).

Note that the three groups of factors listed above do not operate independently of each other but are strongly linked. For instance, environmental variation has a major influence on the outcome of local interaction as competition among individuals is stronger in nutrient-poor locations than in nutrient rich locations (Chesson 2000). Similarly, local interactions may have an influence on the growth of individuals (Schneider et al. 2006).

2.1.4 The mean field approach and individual-based modelling

Historically, the theory of plant population dynamics has taken its origin in animal ecology and until very recently the mean-field assumption, which has its origins in physics (Weiss 1907; Law et al. 2000), was widely accepted in ecology, claiming that interaction may be described on the basis of the average abundance of different species across space. However, the assumption only holds if species are highly mobile, organisms are strongly mixed and the interaction takes place over long distances. Whereas this might hold for some animal populations, it is certainly not true for many plant communities.

Plants are static. The individuals' immobility and the prevalent spatial structure make it impossible for organisms to encounter each other in proportion to their average density (Law et al. 2003). Furthermore, organisms are typically not strongly mixed but often occur in patches or are locally clustered (Wu and Levin 1994). As a result, most interactions take place over very short distances only (Tilman 1994). Thus, individuals may only compete for resources within their own area of interaction even though ample resources are available at larger distance. Therefore, the local density expe-

rienced by individuals is substantially different from the overall population density (Law et al. 2000; Murrell et al. 2001). Competition takes place in space in a very restricted area of a size only a few times the size of the individual plant (Bolker et al. 2003) and individuals experience only that which occurs in their immediate neighbourhood (Tilman 1994).

Consequently, the mean field assumption is unlikely to be valid in the context of plant communities. Ecologists have therefore moved away from the mean-field approach and ecological research has become more interested in taking on what has been called the "plant's-eye view" (Purves and Law 2002). As a result, spatially explicit individual-based modelling has become a very popular approach in ecological modelling over the last ten years (Winkler et al. 1999; Ermentrout and Edelstein-Keshet 1993; Law et al. 2003).

2.1.5 Individual based modelling and the plant's eye view

"This is the age of the individual-based, spatially explicit, computer-based model" (Law et al. 2000).

Over the last decade ecologists have become more and more aware of the necessity to model ecological plant communities from an individual perspective and have started using individual-based models (IBMs). These model the behaviour of each member of a biological population as an individual (DeAngelis and Gross 1992; Huston et al. 1988; Judson 1994). They are mechanistic models built on equations capturing those properties of the individuals which are believed to influence community dynamics. IBMs are based on two fundamental principles:

a) The individual organisms are considered behaviourally and physiologically distinct entities and;

b) interactions among individuals are inherently localised, i.e. organisms are only influenced by the individuals in their vicinity.

It is generally accepted that in considering intra-species diversity the result of different genetic and environmental influences on each individual is fundamental when structure and functioning of ecological communities are modelled (Loreau et al. 2001). However, most IBM models do not consider inter-individual differences within a species, i.e. do not take diversity below the species level into account. A recent approach described in Bown et al. (2005) and Pachepsky et al. (2005) has tried to resolve this shortcoming.

Individual-based simulations yield realisations of complex deterministic and/or stochastic processes. Observing simulated systems that evolve over time reveals the behaviour of the system as well as emerging phenomena. The behaviour of the systems may be analysed in relation to the parameters of the model and the behaviour is then interpreted with reference to the assumed underlying ecological processes. This allows an assessment of the behaviour of the system under many different conditions over extended periods of time – something that could never be done in an experimental setting, due to time and cost restrictions. Furthermore, it generates ideas and hypotheses.

In general, the models are very complex, and based on large numbers of equations. Law et al. (2000) present a hypothetical example where a community modelled on a spatial lattice consisting of 100 cells and containing 10 different species already requires 1000 equations. The authors comment

that this is too complex and suggest that simplified models more amenable to interpretation are necessary, but state that appropriate methods that are capable of reducing the complexity were currently not available.

2.1.6 A case for spatial statistical modelling

Law et al. (2000)

> "Ecology needs new ideas and methods to deal with dynamics of processes in a spatial setting."

The authors make a case for projecting the complex dynamics into lower-dimensional space as it is clearly not the location and properties of each individual that we are interested in but more general properties of the plant community. It is not clear, though, how the dimensionality of the space projected into should be chosen. On the whole, a lower dimensional projection of the system would yield a less detailed description of the observed pattern focusing on its overall properties, which may be interpreted more readily. The individual-based models in their current form yield simulated realisations, based on the properties of individuals but do not capture overall properties of the entire community and little work has been done so far that attempts to reduce the system's complexity (Bown et al. 2005; Pachepsky et al. 2001).

Spatial patterns observed in ecology are the result of a very large number of underlying mechanisms and processes, many of which are difficult to measure *in situ* and the relative importance of each process can be context dependent. A model that tries to describe these underlying mechanisms attempts to explain the main structures in the community by focusing on

only a small number of mechanisms. These govern the overall characteristics of the pattern formed by the community, i.e. the ecological signal in which ecologists are most interested. The remaining mechanisms might each influence the pattern in their own small way but may not be of general interest. An individual-based model tries to include deterministically as many of the detailed mechanisms involved as possible. However, the model does not acknowledge that some of these mechanisms might be less important and, what is more, that some important mechanisms have not been incorporated in the model. The latter may be addressed by including stochasticity into a model, i.e. a random mechanism that mimics those mechanisms in a system that are not explicitly accounted for. This has been successfully done (Pacala et al. 1996), but still it is neither possible to distinguish between or quantify the effects of the random component on the outcome nor to quantify the relative importance of the assumed underlying mechanisms (Law et al. 2000). As a consequence, it is not possible to decide for an IBM model whether the observed phenomena have been caused by those mechanisms most relevant to the natural system or whether they have merely been caused by the random component, i.e. by those mechanisms that have not been explained by the model.

As noted, underlying ecological processes and the resulting patterns are intimately linked, even if the link is not unique or isomorphic. Therefore, changes in the properties of the process will have an influence on the properties of the spatial pattern. In order to understand fully this relation, it is desirable to predict properties of the simulation, i.e. a realisation of a model from the properties of the model. However, being neither directly linked to

the resulting spatial pattern nor yielding any mechanism for describing the spatial properties of the resulting pattern, individual-based models do not provide a technique of predicting resulting patterns from the properties of the model, i.e. the parameters.

As a consequence of the above points, we suggest the use of spatial statistical models in the context of modelling ecological communities. These models are based on observed spatial patterns but incorporate background knowledge on the presumed underlying processes, if available. For instance, if knowledge on potential intra- and inter-species interaction and an interaction between individuals and the environment is available, this may be explicitly incorporated into the model. Statistical models are constructed such that a small number of initially unknown parameters describe the overall properties of the system. These parameters are related to the spatial properties of the observed pattern. After an initial analysis of these properties a model is chosen that is capable of accounting for the observed spatial properties and the parameters are estimated from observed data, i.e. a spatial pattern. Consequently, the estimated parameters of the model may be meaningfully interpreted and their relative importance, i.e. the relative importance of the mechanisms that the parameters reflect may be determined.

The models contain one or several random terms, which account for those mechanisms that are unobservable or unknown as well as for random noise (Cox et al. 2000). Since the models are statistical, their goodness of fit to the data may be assessed. Furthermore, the amount of variation in the data explained by the model may be established in order to assess how much the assumed underlying processes have contributed to the formation of the

specific pattern and how much of the information contained in the pattern is not captured by the model. In addition, the models can be used to make predictions relative to changes in the environmental conditions or within the model.

2.2 Existing spatial statistical methods for ecological data

Spatial statistical techniques include many different methods that aim to quantify spatial patterns or incorporate spatial information into a model. Many of these methods have been developed independently, in different areas of science (Dale et al. 2002), before they were applied to ecological data (Liebhold and Gurevitch 2002; Hoef and Cressie 1996).

On the one hand, the large number of methods available reflects the diversity of data (Dale et al. 2002), but it is this same range of approaches that has very often impeded communication between statisticians and ecologists (Liebhold and Gurevitch 2002). Ecologists have often been confused by the multiplicity of available techniques, and mystified by their interpretation (Perry et al. 2002). Furthermore, the "languages" used by ecologists and statisticians differ substantially (Dieckmann et al. 2000), in particular in view of statistical methodologies becoming increasingly sophisticated (Dale et al. 2002) and it has been felt that formal mathematical language is not as intuitively understandable as the algorithms used in computer-based models (Perry et al. 2002).

In most cases, spatial statistical methods have been used as a purely

empirical approach, and were thus rarely model-based but merely characterising or summarising data (Perry et al. 2002). As a consequence, ecologists have often not been aware that statistical models are available that may be applied to ecology, and this lack of awareness may impede the development of ecological theory.

Traditionally, the subdivision of available methodologies in spatial statistics has followed the taxonomy established by Cressie (1991) although it has not always found universal support (Perry et al. 2002). We use a slightly amended version of the subdivision here to aid explanation. Our subdivision is based on the types of data to which the statistical models are fitted. More importantly, we also base it on the types of questions that may be addressed with a specific set of methods. Generally speaking, all these methods attempt to model one or several variables (the *outcome variables*) on the basis of other variables (the *explanatory variables*). Basing the taxonomy on the type of questions that may be addressed ultimately implies basing the taxonomy on the type of the outcome variable of interest. More specifically, different methods have to be applied depending on whether the outcome is a spatial variable, i.e. the location of individuals or the spatial pattern itself, or a non-spatial variable that is subject to an underlying spatial dependence or spatial autocorrelation. In this section, we describe briefly three main groups of spatial statistical methods based on this taxonomy. In the first two methods (Area referenced data and Geostatistics) the outcome is a non-spatial variable. The third method (spatial point process modelling) has a spatial variable as outcome variable. We show that for our purposes spatial point process models are the most appropriate method, and the other types

of spatial statistical methods help contextualise the selected method.

2.2.1 Non-spatial variable as outcome

Traditional, i.e. non-spatial, statistical methods assume that data values are independent of each other, but this assumption is often violated in an ecological context (Liebhold and Gurevitch 2002) such that standard methods are not valid. Hence, methods have been developed which resolve this problem. The methods described in this section typically commence the modelling process by describing the spatial dependence and consequently model the variable(s) of interest based on the estimated dependence structure.

Area referenced data Area referenced data are frequently referred to as "lattice data" (Cressie 1991). These consist of measurements taken on a regular or irregular lattice where information on the neighbourhood structure within the lattice is given. The methods that deal with area referenced data may be considered a generalisation of time series analysis (Upton and Fingleton 1988). Whereas in time series analysis we have measurements taken over time with a one-dimensional dependence among measurements in two directions, the spatial analogue has an at least two-dimensional dependence structure in infinitely many directions. The aim is to describe the behaviour of the non-spatial variable of interest and relate it to covariates or the spatial location itself.

Geostatistics The area of statistics conventionally called geostatistics has its origin in mining applications and deals with quantities which are subject

to continuous spatial variation, such as nutrient contents in the soil (Jackson and Caldwell 1993). However, the quantities cannot be observed continuously but are measured at discrete locations (Diggle et al. 2003). Geostatistical methods seek to infer and predict the properties of the quantities of interest over continuous space from these measurements. Geostatistical methods have been successfully applied within ecology (see Rossi et al. (1992) and Liebhold et al. (1993)).

2.2.2 Spatial variable as outcome

The methods described in this section are concerned with modelling the spatial pattern itself, based on explanatory variables, if available.

Spatial point process modelling Spatial point processes are stochastic models that describe the spatial pattern formed by the locations of objects (e.g. plants, houses, cells) to yield an understanding of the underlying processes that produced the pattern. Hence, the locations of these objects or, more specifically, the pattern formed by these locations (commonly called "points"), is the outcome variable of interest. We introduce spatial point processes in more detail in Chapter 3. For a general introduction to spatial point pattern modelling refer for example to Cox and Isham (1980), Diggle (1983, 2003), Cressie (1991), Møller and Waagepetersen (2003b) and Stoyan et al. (1995).

The data analysed with spatial point process methods typically consist of the x- and y- coordinates describing the spatial location of the objects, i.e. the plants in the current context. Sometimes additional data have been recorded

describing the properties of the objects; these additional properties are either qualitative covariates such as "species", "age-category" of each individual, or quantitative covariates, such as the size of the individuals, etc. Figure 2.2 shows the spatial point pattern formed by the locations of individuals of the species *Dasypogon bromeliifolius* in the data set described in Section 2.4 (Armstrong 1991). The pattern shows an apparent aggregation of the individuals, which might be the result of clonal growth or seed dispersal. Spatial point process methodology would assess this aggregation and fit a model to the pattern that is suitable for describing aggregation (or clustering) as a result of intra-species interaction. If data on environmental variables such as soil nutrient levels were available these could be incorporated into the model as explanatory variables. The relative strength of the influence of both inter-individual interaction and the influence of the interaction among individuals and the environmental variables may be assessed. PhD

In general, spatial point process methodology initially classifies the observed pattern as either random, regular or clustered and investigates spatial homogeneity. It then fits a specific model to the pattern based on this classification and background information on potential underlying ecological processes, such as intra- and inter-specific interactions and interaction with local environmental conditions, as illustrated above. The methodology allows the quantification of the relative importance of hypothesised relations among ecological processes and the resulting spatial pattern.

We believe that spatial point process methods are an appropriate tool for analysing the relationship between underlying ecological processes and the

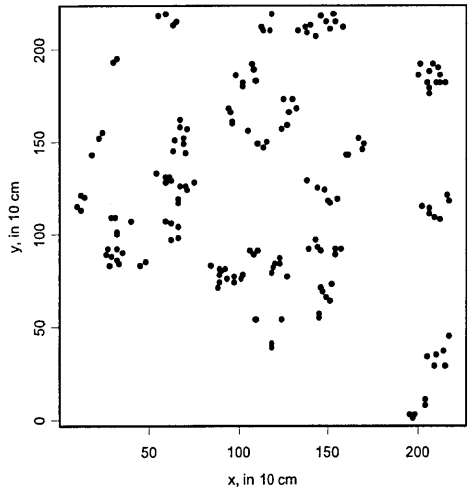

Figure 2.2: The spatial point pattern formed by the species *Dasypogon bromeliifolius* in the data set described in Section 2.4.

resulting spatial pattern. Being based on individual plants, the approach takes the plant's eye-view into account and summarises the observed pattern with a few parameters whilst still taking the individuals as well as the underlying ecological processes into account. Furthermore, predictions can be made, the goodness of fit of the model to the pattern can be assessed and the relative importance of the various aspects of the underlying processes contained in the parameters can be estimated. We will provide a detailed overview of spatial point process methodology in Chapter 3[2].

[2]Note that it is possible to transform point-referenced data into area-referenced data by admitting a lattice on the area and counting the number of points within cells of this lattice. As a result, the number of points in each cell is now the outcome variable of interest. However, this results in a loss of information.

2.3 Modelling the spatial pattern of plant communities

At the beginning of this chapter we established the importance of understanding the factors that promote biodiversity. So far, we have only discussed the importance of considering spatial structure in ecology, emphasised the usefulness of statistical methods in this context and established that spatial point process models may be the most suitable approach. However, in order to understand the impact of human influence on ecosystems we need to consider and model whole plant communities[3]. Whereas some phenomena in ecology might be best analysed for individual species or pairs of species, a number of ecological phenomena cannot be studied without considering entire communities *in situ*. These include selection, competition, mutualism, invasion and succession (Cardinale et al. 2002; Loreau 2000; Loreau et al. 2001; Watkinson 1997). Understanding these phenomena provides a conceptual underpinning for the more general concepts of biodiversity and ecosystem functioning. Hence, community dynamics promoting biodiversity cannot be understood by merely modelling individual species or pairs of species.

In order to appreciate fully the potential contribution of results from a spatial point process analysis on the understanding of community dynamics and theories of biodiversity in general, we need to engage with current theories and key questions that relate to these. Section 2.3.1 outlines briefly the theoretical background necessary for this; Section 2.3.2 introduces contribu-

[3]Following Crawley (1997b) we define a plant community as all the plants occupying the area being studied.

tions that spatial point process modelling could make to the development of ecological theories on biodiversity.

It is also important that we assess whether the spatial point process methodology currently available is suitable in this context. So far, spatial point process models have not been applied to a complete plant community but only to selected species, at most two or three such as in Mateu et al. (1998) and Diggle (1983). In Section 2.3.3, the consequences of this lack of application and the limitations of the current methodology are outlined.

2.3.1 Ecosystem functioning, biodiversity and species coexistence

As a consequence of human interference with the natural world recent decades have seen an unprecedented loss of biodiversity with an extinction rate for mammals in the last 400 years that has been estimated to be 50-100 times higher than the rate estimated from fossil records (Regan et al. 2001). This has resulted in a growing concern about the impact of this on the future development of ecosystems and their functioning (Mouquet et al. 2002). This increased extinction rate is mainly due to habitat loss resulting from a number of influences, e.g. harvesting, exotic species that become dominant and climate change (Armsworth et al. 2004). An increased extinction rate clearly reflects some of the aspects of our understanding of loss of biodiversity but does not cover the concept in its entirety; the following paragraphs yield a more refined definition of the concept of biodiversity, and introduce and discuss current biodiversity theories.

2.3.1.1 Defining and quantifying biodiversity

The term "biodiversity" was coined by W. Rosen in 1982 (Wilson 1988) by contracting the words "biological" and "diversity". But biodiversity is not to be interpreted as merely a large number of different species, as conservation could then be effected via zoos or other artificial environments, but also refers to the variation of life forms in their natural habitat with a range of biotic and abiotic ecological and evolutionary interactions. Biodiversity refers to variation on all levels of biological organisation, from genes and chromosomes to communities of ecosystems (Crawley 1997a; Armsworth et al. 2004).
A number of different concrete definitions of biodiversity have been suggested and to date researchers have not agreed on a single term of reference (Noss 1990). For instance, the UN Convention on Biological Diversity defines biodiversity as

> "(...) the variability among living organisms from all sources including, *inter alia*, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems" (United Nations Environment Programme 1992)

We use this definition here. There has been a tendency to focus on the species level of biodiversity. This thesis will concentrate on this, as well, acknowledging that considering biodiversity at finer and coarser levels is equally important (Armsworth et al. 2004).

Two concepts have been central to the quantification of biodiversity:

- *Species richness:* the number of species in a given area and

- *Species evenness (equitability):* the relative abundance with which each species is represented in an area.

Thus, an ecosystem where all the species are equally abundant has high species evenness. An ecosystem where some species are represented by many individuals, and other species are represented by very few individuals, e.g. ecosystems with log-normal species abundance curve frequently found in natural systems (Pachepsky et al. 2001), has a low species evenness.

The most commonly used index for quantifying diversity is the Shannon-Wiener diversity index (e.g. Greig-Smith 1983), which has its origins in information theory and is defined as follows:

$$H' = -\sum_{i=1}^{S} p_i \ln(p_i),$$

where S is the total number of species at a site and p_i is the relative abundance of species i. The index increases with S and is maximised for a given value of S when all the p_i terms are equal, i.e. the index incorporates both richness and evenness[4].

Petchey et al. (2004) argue that the extent of functional differences among species has a strong impact on ecosystem processes and hence criticise that species richness assumes that all species are equally different. They emphasise the importance of including functional diversity, in terms of the distances

[4]There are other approaches to measuring biodiversity which also take evolutionary diversity into account, by considering the evolutionary relatedness of the species present in an area, see, e.g. Faith (1994). These will not be considered in this thesis. Similarly, (Petchey et al. 2004) suggest a measure of functional diversity. We acknowledge its relevance in particular in the context of ecosystem functioning but are unable to apply it in this thesis, due to a lack of appropriate data.

among species in the trait space formed by functionally important species' traits.

Different spatial scales of biodiversity are typically distinguished (modified from Armsworth et al. (2004)):

- α *diversity:* the local diversity within each site.
- β *diversity:* the change in species composition from one site to another.
- γ *diversity:* The diversity measured over the entire suite of sites being considered.

In this thesis, we are mainly concerned with the analysis of a plant community with high α-biodiversity, even though we acknowledge that biodiversity on different scales is not necessarily independent. In fact, the general area around the actual study site has been reported as being biodiverse on all three scales (Dixon 2005).

2.3.1.2 Models of biodiversity

Most biodiversity studies consider species abundance and results typically show that in natural communities only a very small fraction of the species represents the majority of the organisms and many species are represented by a few individuals only (Whittaker 1972; Pachepsky et al. 2001). This observed structure is sought to be explained by ecological theories described in the following paragraphs.

Species richness and the competitive exclusion principle Classical ecological theory has maintained that species that share the exact same

environment or habitat rely on the exact same resources. Hence, different species would be unable to coexist if they exploited these resources in exactly the same way, since the best competitor would outcompete all other species (de Mazancourt 2001; Crawley 1997b). Thus, according to the competitive exclusion principle (or *Gause's law*) two species cannot coexist if they compete for the exact same resources (Gause 1934). However, this principle only holds when the following conditions are met (adapted from Palmer (1994) and Wright (2002)):

1. The community is sufficiently established
2. The environment has no temporal variation
3. The environment has no spatial variation
4. Growth is limited by a single resource
5. Rarer species are not favoured in terms of survivorship, reproduction, or growth
6. Species have the opportunity to compete
7. There is no immigration.

However, in the natural world these conditions do not hold. The environment, for instance, is typically neither temporally nor spatially constant, violating conditions 2. and 3., leading to an expectation of an increased species richness with larger environmental variability. In addition, there is evidence that species have a tendency to be specialists in the way they survive, reproduce and grow, violating condition 5. Hence, a larger number

of coexisting species should be expected if species adapt in very different ways to a particular environment. *Palmer's corollary* postulates that the more these conditions are broken the higher the number of species that can coexist (Palmer 1994).

All this indicates that coexistence can only be maintained if species use the resources in different ways, i.e. adapt differently to the environment. In other words, coexistence relies on inter-specific trade-offs with regard to species' properties, for example between root and reproductive allocation (Tilman 1994). This has led to the development of niche theory.

Niches and niche-structured communities The term ecological niche, which originates in Elton (1927) and Grinnell (1917) and more formally in Hutchinson (1957, 1959), refers to the multidimensional summary of a species' ecological attributes, including its abiotic tolerances, its maximum relative growth rate, its phenology, its susceptibility to enemies and its relative ability to compete with other plant species (Crawley 1997b). Hutchinson (1957) suggests a niche hypervolume for each species rendering ecological communities highly complex.

We distinguish between the *fundamental niche* and the *realised niche* of a species: Fundamental niches are those that the species occupy in the absence of competition. The realised niche is that which the species occupies where competitors exist; there has been strong evidence that the two types of niches are not necessarily the same (Hutchinson 1957; Crawley 1997b).

Classically, plant ecological theory has maintained that different species occupy different niches, i.e. that they have adapted to survive in a specialised

environment by exploiting the available resources. Speciation prevents inter-specific competition from driving competitive exclusion and thus promotes long-term coexistence of several species in the same area. As a result, the theory predicts that inter-species competition is rather low.

Niche theory states that species-specific differences influence the population dynamics and therefore the behaviour of the whole community (Purves and Pacala 2005). As a consequence, biodiversity and functioning are directly linked from a niche theoretical perspective; loss in biodiversity directly influences a community's functioning.

Random drift and the unified neutral theory of biodiversity and biogeography The theory of niche specialisation was challenged in the 1970s by a number of authors (May 1975; Caswell 1976) and by Hubbell in particular (Hubbell 1979). These authors demonstrated that phenomena typically found in natural communities such as the distribution of population sizes and the log-normal species abundance curves could be produced by very simple neutral models (Chave 2004). Bell (2001) termed these models *neutral community models*, the most famous of which is Hubbell's "Unified Theory of Biodiversity and Biogeography" (Hubbell 2001).

Hubbell's theory, often also referred to as *neutral theory* or *random drift theory*, has been derived from MacArthur's and Wilson's *theory of island biography* (MacArthur and Wilson 2001) and is an analogue of genetic drift (Gaston and Crown 2005). It states that all species and thus all individuals in a community are equivalent, such that they are interchangeable, independent of environmental conditions and of space and time (Purves and Pacala 2005;

Chave 2004). The theory acknowledges that while there are differences in species' properties, these do not affect the population dynamics and hence have no impact on the behaviour of a community or its biodiversity (Hubbell 2001). In fact, in a completely neutral community all but one species can be eliminated without affecting the biochemical functioning of the community.

The theory postulates that species abundances follow a random walk or *drift*, with equal per capita probabilities of birth and death. The only constraint here is that the total number of individuals over all species in the community is constant (de Mazancourt 2001). There is no superior competitor and the probabilities of death and birth, dispersion patterns and speciation are equivalent for all species. The community is saturated with individuals; as soon as one individual dies its place will be taken up by another individual, leading to strong competition (Gaston and Crown 2005).

Niche structure or ecological drift? It is not clear whether it is the formation of niches due to differences among species, or ecological drift due to species' equality that has the greater impact on biodiversity. There has been evidence in favour of both theories (Bell 2001; Hubbell 2001; Chave 2004; Condit et al. 2002; Duivenvoorden et al. 2002).

The niche theory is supported by the fact that characteristics of vegetation are clearly linked to climate (Walter 1973, 2002; Archibold 1995), i.e. there is no neutrality at a global scale (Purves and Pacala 2005). Furthermore, it has been shown that species vary along a number of trade-offs, including a) growth-rate in high light versus survival in low light (Bazzaz 1996; Pacala et al. 1996), b) physical segregation of species into defined habitats (Harms

et al. 2001; Phillips et al. 2003), c) competition versus colonization abilities (Levins and Culver 1971) d) the varying ability of species to resist specific pests (Janzen 1970; Connell 1971).

On the other hand, there is also evidence in favour of the neutral theory. The fact that the distribution of population size within a community tends to follow log-series of log-normal distributions for very different types of communities (Hubbell 2001) suggests that every detail of the species properties in the communities cannot have a strong impact on the community structure but that more general processes operate. Hubbell (2001) and Bell (2001) show that models derived from neutral theory can reproduce typical patterns observed in community ecology, such as features of the species-area curve, the distribution of range sizes, the range-abundance relationship, and turnover in community structure in space.

The distinction between the two theories and the testing of the validity of either of them is crucial. If one were able to prove that the neutral theory is valid rather than the niche theory the conclusion would be that biodiversity and ecosystem functioning are not linked. This, in turn, would imply that the reduction of biodiversity does not reduce the fitness of an ecosystem and the elimination of one or more species will not have an effect on the ecosystem (Purves and Pacala 2005).

Purves and Pacala (2005) try to resolve this problem by formally demonstrating that the two extremes are not contradictory, by introducing niche structure into a theoretical neutral model. The results show that this introduction does not influence the species abundance prediction derived from the neutral theory. They conclude that:

> "In diverse communities, the distribution of species abundances is determined entirely by ecological drift, independent of niche structure; but the biogeochemical functioning is determined entirely by niche structure, independent of drift."

Similarly, Gaston and Crown (2005) argue that there might be some truth in both approaches. Chave (2004) even goes as far as claiming that the two theories are not contradictory or conflicting but are complementary, as niche theory mainly explains coexistence in species-poor communities governed by purely deterministic processes and fixed rules whereas neutral theory applies to species-rich communities where stochasticity has to be included to account for the survival of many rare species.

2.3.2 Potential contribution of spatial point process modelling to open questions

In an attempt to provide evidence for or against the random drift theory, Stephen Hubbell and his colleagues established a 50 ha plot on Barro Colorado Island (BCI) in Panama in the early 1980s recording the locations and sizes (diameter at breast height) of 235,349 individuals of 304 (rainforest) tree species in 1982 (Condit et al. 2000; Burslem and Law 2005). In addition, a large number of soil variables was collected. Since then, similar data have been repeatedly collected at regular times (Condit et al. 2002). Similar plots have been established in a network of 16 forest plots in several countries coordinated through Center for Tropical Forest Science (CTFS) of the Smithsonian Tropical Research Institute in Panama (http://www.ctfs.si.edu). A large number of biodiversity studies has been conducted using the data col-

lected from these plots but to date spatial point process methodology has not been applied to the data sets, other than in a descriptive way (Coomes et al. 1999). We envisage that spatial point process modelling may contribute to the discussion and outline potential points where the methodology may be applied in the following paragraphs.

Interaction strength and neighbourhood identity The fact that these data sets have been collected on a spatially explicit basis provides a platform upon which an understanding of the inter- and intra-specific interaction structure in the community, which takes place at a very local scale, may be developed. Neutral theory predicts identical competitive ability of all species, whereas niche theory predicts that an individual is differently influenced by individuals from different species (Uriate et al. 2004, 2006). This would imply that species have a varying strength and range of interaction.

Evidence of strong inter- and intra-specific competition would provide evidence in favour of the neutral theory. However, low inter- and intra-species interaction as well as strong variation in strength and direction of inter- and intra-species interaction between species would provide evidence in favour of niche theory (Uriate et al. 2004, 2006). Spatial point process models typically include parameters representing different types of interaction and their strength, direction, range and significance may be assessed (van Lieshout 2000), i.e. the models may be applied to inform on these aspects of community dynamics.

Microhabitat specialisation Evidence of microhabitat specialisation would provide arguments supporting the niche theory as this would be an indication of species being not equal but adapted to a specific environment (Harms et al. 2001). Niche theory predicts that species have adapted to specific environmental conditions and, under this hypothesis, are expected to be more likely to be found in specific microhabitats and thus to be aggregated within these (Burslem and Law 2005). A spatial point process analysis could detect aggregation and could model the spatial pattern formed by the individuals taking local environmental conditions into account. By merely analysing the spatial pattern formed by species alone it is not possible to distinguish between aggregation due to interaction or environmental heterogeneity unless data on the environmental conditions in specific locations are available.

Density dependent processes Niche theory is linked with the *Janzen-Connell hypothesis* (Janzen 1970; Connell 1971) which predicts that as a result of speciation mortality of plant seeds and seedlings is negatively related to distance from conspecific adults. In other words, the further away a seed or seedling is from an adult plant the less susceptible it is to pests and the better its survival.

This indicates that locally rare species have a recruitment advantage and the hypothesis predicts that there is an increase in inter-tree distances with increasing tree size making adult trees less aggregated than juveniles. An appropriate model would seek to analyse the effects of habitat specialisation and density dependent interaction.

In this thesis, we analyse a data set collected in the species rich areas of Western Australia that is similar to the data derived from the plots established by Hubbell and his colleagues in the rainforests on Barro Colorado Island (BCI) and the other plots in the CTFS network (see Section 2.3.2) in as far as species richness and spatial explicitness is concerned (see Armstrong (1991) and Section 2.4 for details). It has a high degree of biodiversity but has been collected on a much smaller plot, such that the total number of species is lower. The only data that are available are the locations of the plants. Even though this decreases the complexity of this data set, it is still extremely complex in comparison to the data sets that have been analysed with spatial statistical methods so far where at most two or three species were analysed (Mateu et al. 1998; Diggle 2003).

As indicated, the data set only consists of the spatial locations of individuals but does not provide any information on soil variables or properties of the individuals such as their height etc. We are thus unable to draw conclusions as to the association of species and particular microhabitats at a local scale. However, in the case of this particular dataset it is unlikely that the spatial pattern has been influenced by properties of the soil as this environment can be considered homogeneous at the scale of data collection (Dixon 2005). It will thus enable us to analyse the effects of inter- and intra-species interaction in isolation. Niche theory predicts that, due to speciation, competition in a community should be low but neutral theory predicts strong competition. Thus analysing the interaction strength will provide evidence in favour of or against the existence of species interaction and eventually in favour of either of the two opposing theories of species diversity for the study

data set. In order to do so it is necessary to clarify what exactly is meant be low or strong interaction, see Chapter 6.

The data set's reduced complexity as opposed to the CTFS data makes it easier to handle in the first instance allowing us to explore and develop a suitable methodology. This might yield an indication as to how to proceed in the more complex situation of the CTFS rainforest data sets. We expect that it will be possible to generalise the experience acquired from working with the current data set to larger and more complex data sets, in particular the CTFS plots.

2.3.3 Limitations of available spatial statistics methods

Due to a growing awareness of the role of space in ecological modelling, it is likely that an increasing amount of data detailing individual plants' locations will become available (Burslem et al. 2001) yielding the opportunity for a detailed study of the underlying ecological processes that are driving the patterns. Furthermore, continuously improving technology, e.g. that of geographical information systems (GIS) and increasing computer power, will both support production of more datasets detailing the exact locations of species and facilitate their analysis. Nevertheless, Dale (1999) notes:

> "We have found no example of truly multispecies analysis, in the sense of looking at a combination of species simultaneously, using mapped point data."

So far methods from spatial point process theory have been used mainly in a descriptive way in ecology, typically using summary statistics such as Ripley's

K-function (Perry et al. 2002). However, descriptive statistical methods only form the first step in the analysis of a spatial pattern and describe some of the spatial properties of a pattern. Point process models which would allow more detailed interpretations, predictions and assessment of suitability of the model have rarely been fitted to data (Batista and Maguire 1998).
There have been only a few applications, where spatial point process methods were actually used to model species' patterning, see e.g. Mateu et al. (1998); Stoyan and Penttinen (2000). However, these have only ever taken a small number of species into account; even work by Batista and Maguire (1998), Mateu et al. (1998) and Wiegand and Moloney (2004) are restricted to analysing the patterns of a single species or pairs of species but do not take the whole community into account. Only very recent work has considered entire plant communities (Illian et al. 2004, Illian et al. 2005).

This shortcoming may be partly due to a lack in communication between statisticians and ecologists leading to an unawareness of methodology on the part of ecologists but also to the fact that the available methodology is not yet appropriate for application to the complex needs of ecological modelling (Liebhold and Gurevitch 2002). Furthermore, ecological modelling has used mainly non-statistical mathematical models and techniques. Statistical models have often not been considered by scientists mainly trained using these non-statistical approaches and even if they have been considered their full potential has not been recognized or understood, as for instance in Neuhauser (2001).

The pattern of a plant community may be very complex due to the large number of species and the resulting extremely large number of potential

inter- and intra-species interactions. The complexity of a spatial point process model increases at best linearly but often exponentially with the number of types in the multi-type point pattern sought to be modelled. This is due to an increasing number of parameters complicating model simulation as well as parameter estimation eventually leading to difficulties in the interpretation of results. However, not only may the actual fitting of the spatial point process model be difficult due to a lack of methodology and the inherent complexity of the problem, but also the initial explorative inspection using summary statistics is already complex. In addition, no methodology exists that characterises the spatial behaviour of all species simultaneously by describing the main characteristics of spatial behaviour taking the spatial pattern of all species into account.

In other words, so far, no methodology exists to

- parsimoniously apply explorative statistics to a complex point process data set;
- reduce the dimensionality of the data set by concentrating on the main aspects of spatial behaviour;
- parsimoniously fit an appropriate, highly multi-species point process model to the data set.

This thesis aims to address these shortcomings by developing an appropriate methodology and analysing a case study data set (see Section 2.4 for details) that in many ways resembles the CTFS plot data. However, being

less detailed and restricted to a smaller area with homogeneous environmental conditions, it is easier to develop methodology, which may eventually be generalised to be suitable for more complex data sets such as those collected by Hubbell and his colleagues.

2.4 An example of a biodiverse plant community

In this thesis, we consider a multi-type spatial point pattern formed by a natural plant community in the Mediterranean type shrub and heathland of the southwestern area of Western Australia (Beard 1984). The study site, located within the Tiwest Joint Venture mineral sands mining lease at Cooljarloo, is approximately 18 km north of the Cataby Road House on the Brand Highway and 150 km north of Perth. The data are the locations of 6378 plants from 67 species on a 22 m by 22 m plot (Armstrong 1991). For simplicity, the data set will be called the "Cooljarloo data set" from now on.

Lying within a mineral sand mining area, the study area was mined shortly after data collection in 1990 and will have to be rehabilitated as soon as mining has ceased. Current efforts of rehabilitation in neighbouring areas, however, have resulted in a very small survival rate for some species. Rehabilitation, nevertheless, is a legal requirement after large scale mining.

This thesis analyses and models a large part of the patterns of the 67 species. Nevertheless, not all species have been taken into account, typically as a result of low abundance. In Chapters 4, 5 and 6 we give details as to

Figure 2.3: Photograph of the vegetation of Banksia dominated shrub and heathland in an area similar to the study site

why the particular species have been analysed with the respective methods and why others have not been considered. Table 2.1 below lists the species contained in the dataset that have been analysed and modelled in this thesis. Refer to Appendix A for a list of all species in the data set.

2.4.1 Data collection methods

This section summarises the data collection methods applied in Armstrong (1991). The initial site was fixed by setting a 30m tape measure out on a N-S compass bearing for a distance of 22m. This line was marked by driving fence droppers (107cm long) into the ground at 1m intervals along the tape measure. A compass was used to set the second side of the plot at right angles to the first side, checking accuracy with a 6m to 8m to 10m triangle.

Again, fence droppers marked the line. Setting tape measures on the ends of the first two lines, the third and fourth lines were set. The final corner of the square was where the two tapes crossed at 22 m. As a check, compass bearings were taken. These were within 1 degree of the required bearing. Again droppers marked the lines.

A 1m by 1m grid was set out using 3.12mm fencing wire, to cover the entire plot. The wires were strung between droppers on opposite sides of the plot. Where tall shrubs obstructed the line, the wire was laid on the ground below the foliage. There were no large trunks on these lines. On completion of laying out the wire grid, distances were checked for accuracy along both north-south and east-west sets of lines (1.0m within 2cm) and adjusted as required.

Trampling was kept to a minimum by walking between plants wherever possible and only walking in the plot to place the fencing wire. All other movement was restricted to outside the plot while setting up the study area. Recording of vegetation was undertaken by noting all plants within a 1m by 1m quadrat. The order in which individual quadrats were recorded was from the north west corner (A1), proceeding sequentially east along the north boundary (A1 to W1). The next east-west line was then recorded, (A2didi to W2) moving sequentially south until the final line (A22 to W22) was completed. A record of each individual plant's location and species was made on previously prepared proformae. Each sheet covered four individual 1m by 1m subplots. The location of each plant was visually located to the nearest 10 cm, each plant was recorded in a cell on the proforma. Each cell on the proforma represented 10cm by 10cm on the ground.

Several species were strongly clonal, for example, *Alexgeorgea nitens* and *Conostylis canescens*. The identification of individual plants was virtually impossible. To overcome similar difficulties Magurran (1988) suggests the adoption of modular units as the base unit. These modular units can be shoots of a tree, a tiller of grass, a clump of a clone or some consistent attribute of the species. Field recordings used these modular units as the element of vegetation, based on a premise that these modules are proportional to niche size.

The accuracy of the visual estimation was checked by measuring the location of each plant in the first 2m x 2m square of the plot. This resulted in three plants out of the sixty-one placed in adjoining cells. The error distances, 2cm, 4cm and 7cm were less than the cell size (10cm by 10cm). At the time of data collection, it was assumed that greater accuracy would not have been necessary for the analysis undertaken in the original study (Armstrong 1991). Furthermore, it was estimated that the additional time necessary to achieve increased accuracy would have extended the duration of the field work to far longer than was available for the field investigation.

Rather than attempting to write the name of each species in the 10cm x 10cm square on the proforma, numbers were used. These corresponded to the species in the Tiwest Joint Venture's flora database. This list contained all 307 species recorded from the mining lease. A search of this list was made for all species known to occur in Banksia woodland. The resulting list contained 97 species, 67 of which were found within the 22m by 22m study plot.

2.4.2 General characteristics of the southwestern West Australian flora

The flora of the southwest of Western Australia[5] is particularly rich in plant species (Crawley 1997b). Here, a distinguishing feature is the large number of *endemic* species with *small populations*, *restricted distributions* and a general *rarity* of many of the species (Brown and Hopkins 1983; Hopkins et al. 1983; Coates and Atkins 2001). Beard et al. (2001) estimate that 79% of the flora is endemic. This is surprising in an area of low relief, as other regions of the world that also exhibit high numbers of endemic species usually have much higher and more extensive mountain ranges, for example South Africa, California, Turkey and Greece (Hopper 1979).

A considerable amount of investigation has been undertaken within the Southwest of Western Australia in an attempt to explain this high diversity in a region with very little relief (for example Marchant (1973); Hopper (1979); Hopper et al. (1996); Coates (2000)). The proposed causes for this include: long periods of geological isolation (Marchant 1973; Hopper 1979; Hopper et al. 1996); fluctuating climatic conditions (Marchant 1973), a stable ancient landscape that has experienced extensive fragmentation (Hopper et al. 1996), the stress of a transitional climatic zone (Hopper 1979), and very impoverished soils resulting in habitat specialisation (Beard et al. 2001). All these factors have combined to produce a rich, biodiverse flora that is inextricably linked with rarity, endemism and disjoint populations. In fact the Southwest of Australia is one of the world's biodiversity hot-spots (Crawley

[5]See Beard (1990) for the exact location of the southwest of Western Australia.

Figure 2.4: Photograph of the soil in an area similar to the study site; photograph courtesy of Paul Armstrong

1997a). This refers to a high α, β as well as γ diversity but here only the α diversity can be assessed.

2.4.3 Soil properties

The characteristic sandy soil in the area is extremely low in nutrients and water (Armstrong 1991) making it similar to crushed glass (Armstrong 2005). See the picture in Figure 2.4 as an illustration of this.

More specifically, the study site is located on a soil type referred to as Bassedean Dune System (Beard 1990). It has a subdued topography with swamps in the swails of the Pleistocene dunes. These dunes are formed from deep sand derived from quartz with low nutrients and low humus content. The soil is very water porous and free draining (Elkington 1991). The flora

has been evolving in the area for thousands of years and has speciated significantly to survive in these nutrient poor soils (Hopper 1979) resulting in a high species-diverse community. In the particular area where the data were collected the water level is very deep, and has been estimated to be at least two metres below the surface (Armstrong 2005).

Whereas the soil is very heterogeneous at a larger regional scale (Brown 1989; Griffin et al. 1983), it can be considered homogeneous at a local scale (Dixon 2005; Armstrong 2005). Perceived clustering at a local scale can be considered a result of inter- and intra-species interaction rather than resulting from variations in soil nutrient and water levels. More details about the soil properties in the specific area may be found in Armstrong (1991).

2.4.4 The role of fire

Fire is a naturally occurring phenomenon in many Australian biota with only a few exceptions in some areas of Tasmania (Bell et al. 1984; Bradstock et al. 1995). The frequency of fire is dependent on the productivity of the biota and the climate. It has been estimated that prior to European settlement, fire occurred at frequencies of typically five to 50 years[6] (Bell 1984; Ward et al. 2001; Enright et al. 2005). Since European settlement this has changed to five to ten years. The period since the last fire in the study area (1980) to the time of the data collection at Cooljarloo was ten years. This would have allowed a sufficient time for a relatively stable community to have developed. Some species referred to as fire ephemerals germinate very quickly after a

[6]Some areas in the Kimberly experience annual fires whereas other central areas have very infrequent fires of 100 or 200 years apart (Enright et al. 2005).

fire and are only found in an area for 2 to 8 years post-fire with seeds stored in the soil seedbank (Brown 1989); these could thus not be considered in this current study. Plants have adapted to the regular fires, through the development of fire regeneration strategies.

2.4.4.1 Fire regeneration strategies

Many species in Mediterranean climate have evolved regeneration strategies to cope with fires. These fires have the potential to remove most of the above ground biomass (Bell et al. 1984; Bond et al. 1984; Macdonald 1985; Bradstock et al. 1995). Regenerating plants in and/or near disturbed areas may re-establish vegetatively, the so-called *resprouters*, or from stored seed, referred to as *seeders*.

Species that rely entirely on regeneration by setting seed are referred to as obligate seeders. Alternatively, an obligate resprouter relies entirely on regeneration from existing rootstock or stems. However some resprouters are referred to as facultative as they do produce some seed, increasing their regeneration opportunities and gaining the advantages of sexual reproduction (Pate and Dixon 1996; Meney et al. 1999).

In the study data set there were 49 resprouter and 14 seeder species. Three species were both seeders and resprouters and 1 species (with 1 individual) was an annual plant, relying on the established seed bank.

Seeders Seeders have evolved techniques that allow seed to survive a fire. These include (Purdie 1977; Gill 1996; Hughes and Westoby 1992):

- Retaining the seed in woody follicles on the plant for several months to years, only releasing seed after death of the branch;
- elaisomes, ant-attracting organs, entice ants to place seed underground away from the destructive heat of a ground level fire;
- heat and/or smoke to break seed dormancy; and
- annual species, relying on wind blown seed for distribution.

Resprouters Similarly, resprouters have evolved many techniques to avoid the damaging effect of fire on delicate plant tissue. Listed below are several examples (Gill 1996):

- Thick fibrous or corky bark that insulates dormant buried buds and growing tissue which are capable of epicormic regrowth;
- resins and essential oils contained in the foliage that ignites and combusts very quickly, the thick bark providing adequate short-term insulation;
- lignotubers and other forms of enlarged root crown provide both an energy storage organ and sites for initiation of epicormic regrowth; and
- underground organs provide energy storage for the plant, typically fleshy root tubers or rhizomes.

Figure 2.5: The fleshy roots of species *Conostephium pendulum*

These various adaptations contribute to the species' storage effect and their regeneration and persistence niches, which are important for species' survival in a fire prone area (Chesson 1986; Higgins et al. 2000; Bond and Midgley 2001). The persistence niche is particularly important in areas of low productivity. As pointed out by Chesson (1985), environmental variability, spatial, temporal and combined, may contribute greatly to a species' survival in a location and to coexistence between species.

Regeneration of plants in southern Australian heathland occurs largely in the immediate post-fire period (Purdie 1977; Lamont et al. 1985). By germinating immediately post-fire, the maximum growth period may be achieved between fires and nutrients available from the ash bed may be utilised (Auld and Tozer 1995). However, some inter-fire seedling recruitment has been recorded, usually associated with senescence, drought or disease. This can

maintain a species presence during extended periods without fire.

For seeder species the inter-fire period is usually more critical than for resprouters (Lamont 1996). From simulation trials on Banksia species Enright et al. (1998a,b) demonstrated the relative importance of the inter-fire period on both seeders and resprouters, with different levels of serotiny (seed retained on the plant for more than one year post maturity - high levels of serotiny are when most seed are retained for long periods). A comparison between seeders and resprouters, both with high serotiny, indicated that resprouters were able to cope with more frequent fires (13 year intervals are optimal) than seeders (16 year intervals are optimal), see Enright et al. (1998a,b).

To overcome the unpredictable length of time between fires and hence germination opportunities, many species have evolved dormancy mechanisms to maintain a viable seedbank over a long period. Not until the appropriate trigger mechanism has been applied is the dormancy broken and the seeds germinate. The more common dormancy breaking mechanisms include moisture, indirect or direct heat, chemical mechanisms and time (Kenny 2000; Morris 2000). Several of these triggers may be involved with the dormancy of any given species (Morris et al. 2000). The germination trigger of a combination of heat and smoke following fire is currently being investigated (Keeley 1991; Brown 1993).

2.4.5 Seed dispersal

Seed dispersal occurs predominantly at two times: First on maturity of the seed and second after the seed pod is damaged, typically by fire. When

comparing reproductive strategies, an obligate seeder produces much more seed than a resprouter. Moreover, seeders seek to maintain maximum genetic diversity of the seed, preferably by outcross pollination (Meney et al. 1997). Conversely, an obligate resprouter is not dependent on seed for persistence; hence little energy is channelled into flower and seed production, but is directed to the storage organs.

Once the seed has been produced and released by the plant the next factor in regeneration is the dispersal factor. The dispersal may be the result of a self-propelled or an external mechanism (Levin 2000). These external dispersal mechanisms are wind (winged or plumed seeds), water (seeds that float) or by animals (particularly ants and birds) (Auld and Denham 1999). Distances involved in this dispersal may be only a few centimetres from the point of release, several metres (ants and wind dispersal), or several kilometres (animals - for example emus). Wind can transport seed a few meters to several kilometres from a plant; for example utilising a wing on a large seed or small plumed seeds respectively (Whelan 1997). Most seeds transported by animals are carried either in the gut (via ingestion) or on the coat or feathers (adherence) (Whelan 1997).

Some resprouters do produce occasional to small numbers of seeds but these are often not viable and only a very small number or even none of these are actually used for reproduction.

2.4.6 Specialised growth behaviour

Trunk thickening Trunk thickening is an adaptation of certain resprouter species (e.g. *Eremaea astrocarpa, Eucalyptus todtiana* and *Hibbertia hyper-*

icoides) with thickening of the trunk at ground level. This area has many dormant buds, which are capable of resprouting after a fire. This adaptation is most noticeable in many species of Eucalyptus which have a growth form of a mallee, i.e. a multi-stemmed tree form. Many small shrubs, e.g. *Astroloma* also have a thickened main stem near ground level which may resprout post fire. These structures store starches which are used to initiate regeneration of the above-ground portions of the plant. These plants tend to be very long-lived and capable of surviving many fires.

Rhizomatous growth forms Rhizomes are a type of underground stem, which may exhibit either horizontal or vertical growth habit. This growth form is common in the *Restionaceae* family, a very common low rush family occurring within the Kwongan, including *Chordifex sinuosus, Dasypogon bromeliifolius* and *Lyginia barbata.* This growth form, like the trunk thickening, is an important strategy for surviving a fire. Many rhizomes grow below the soil surface, such that they avoid the destructive impact of fire. In addition, many species store starches within the cells of the rhizome which are utilised during the post-fire recovery.

A rhizomatous growth form can also facilitate clonal growth by extending the stem outwards from the parent plant. Over time, the clone may become isolated once the connecting rhizome is severed. These clones may cover considerable distances, up to several square kilometres are known.

Proteoid roots Proteoid roots are largely restricted to the Proteacea family where many genera, including *Banksia*, form specialised root systems

specially adapted to absorb nutrients from low-nutrient soils. These roots form very dense mats within the top centimetre or two from the soil surface. These root structures make it difficult for other species to establish themselves in the vicinity of these plants. It is suspected that some *Ericacea (Leucopogon)* species form mycorrhizal relationships with these Banksia species (Cairney and Ashford 2002; Dixon 2005).

In this area Banksia species are the dominant taller shrub/ low tree species. Hence, these Proteoid roots may play an important function in the species patterning of this area.

2.4.7 Patchiness

Within the general area (less than 5 km radius) eight different plant communities were recorded. These are a reflection of different habitats associated with the variation in topography. The communities are repeated wherever that habitat occurred. These plant communities in the different habitats form a mosaic in the general area. The Cooljarloo data formed only a small portion from within one of these habitats. Thus, patchiness can be observed on a larger scale (several hundred metres) but sampling took place within a single community. As a consequence, no large scale patchiness was observed (Armstrong 2005).

2.4.8 Spatial aspects of the data set

Individual plants may interact in various ways by inhibiting each other's growth while competing for the scarce resources (Richardson et al. 1995).

This is likely to reduce the number of plants in close proximity and thus to influence the observed spatial pattern.

In addition, the poor quality of the soil has a strong impact on the spatial pattern formed by the plants, potentially resulting in a heterogeneous formation at the very small scale micro-habitat level, since plants will only be able to survive in a location where sufficient nutrients and water resources are available. Attraction between plants may also occur if conditions (shade, nutrients, water available) in the microhabitat are made more suitable for another species by the presence of another plant.

The specific regeneration method used by a species will have an impact on the pattern formed by the individual plants (Armstrong 1991) and also on the structure of the interaction between species with different regeneration strategies. Since the resprouting plants have been at exactly the same location for a very long time (some of them for hundreds and even thousands of years (Armstrong 1991)), the seeders, which start anew after each fire, do so with the resprouters already present. We thus assume that the growth of reseeding plants will be influenced by the resprouting plants already present in the plot, whereas an influence of the seeders on the resprouting plants is highly unlikely.

Four of the five most abundant seeder species, *Astroloma xerophyllum*, *Andersonia heterophylla*, *Leucopogon conostephioides* and *Leucopogon striatus* are all from the Ericacea family. These are known to have a unique endomycorrhizal association with ericoid mycorrhiza (Read 1996). The mycorrhiza enable the plants to seize highly complex organic resources in the soil that are normally unavailable and very resistant to microbial decay.

This might enable the plants to break through the proteoid roots formed by the *Banksia* species leading to an attractive interaction between the species rather than a repulsive one (Cairney and Ashford 2002; Dixon 2005).

2.4.8.1 Suitability of the data set in the given context

This thesis is concerned with understanding plant species' coexistence in biodiverse communities together with assessing and improving the suitability of spatial point process methodology in this context. Clearly, for this purpose it is necessary to analyse a data set that consists of the locations of individual plants in a biodiverse community, a criterion that is met by the Cooljarloo data. Further, it is derived from a very old community. This implies that the community dynamics are very well-established (Dixon 2005) and may be assumed to have reached a certain level of stability. This allows the assumption that through evolution the community has reached a state that may be considered relatively optimal in terms of the system's ability to maintain biodiversity and hence a model system in the context of analysing mechanisms of biodiversity.

In addition, the soil may be regarded as homogeneous at a large scale, such that large scale variations in soil nutrient and water levels are unlikely to have influenced the pattern formation (Dixon 2005). As a result, patterning is most likely to be almost exclusively due to intra- and inter-specific interaction. This is very useful in the current context where we are attempting to develop new methodology. We are thus able to concentrate on a smaller number of potential influences on the pattern, yielding methodology that afterwards may be generalised to more complex situations where environmental

variables are likely to have an impact and thus need to be included in the model.

Further, the plot lies within an area where mining activities are very common and legislation requires the re-establishment of the communities. Currently, the replanting of an area is done without taking the orginal spatial structure in the community into account (Dixon 2005). Knowledge on community dynamics and in particular on intra- and inter-species interactions may inform replanting strategies and improve both re-establishment and rehabilitation after mining activities.

species	species number	abundance	regeneration method
Alexgeorgea nitens	3	977	resprouter
Allocasuarina humilis	4	2	resprouter
Andersonia heterophylla	5	686	seeder
Astroloma xerophyllum	6	91	seeder
Banksia attenuata	8	26	resprouter
Banksia grandis	10	1	resprouter
Banksia ilicifolia	11	3	resprouter
Banksia menziesii	12	26	resprouter
Boronia ramosa	14	30	seeder
Bossieae eriocarpa	15	103	resprouter
Conospermum crassinervium	18	266	seeder
Conostephium pendulum	19	61	resprouter
Conostylis canscens	13	149	resprouter
Conostylis juncea	20	28	resprouter
Dampiera linearis	22	24	resprouter
Dasypogon bromeliifolius	23	167	resprouter
Eremaea astrocarpa	25	207	resprouter
Eriostemon spicatus	26	65	resprouter/ seeder
Eucalyptus todtiana	27	1	resprouter
Hibbertia crassifolia	32	96	resprouter
Hibbertia hypericoides	33	148	resprouter
Hibbertia sp.	34	134	resprouter
Hypocalymma xantopethalum	36	96	resprouter
Isopogon linearis	37	68	resprouter
Jacksonia floribunda	38	124	resprouter
Chordifex sinuosus	42	154	resprouter
Lepidosperma angustatum	44	22	resprouter
Lepidosperma tenue	45	61	resprouter
Leucopogon conostephioides	47	657	seeder
Leucopogon striatus	48	251	seeder
Lomandra sp.	49	304	resprouter
Lyginia barbata	50	299	resprouter
Melaleuca scabra	51	377	resprouter/ seeder
Patersonia occidentalis	54	79	resprouter
Phlebocarya philifolia	57	207	resprouter
Platysace juncea	59	22	resprouter
Scholtzia involucrata	61	170	resprouter
Stylidium crossocephalum	64	27	seeder

Table 2.1: List of plant species in the Cooljarloo data set that were analysed in this thesis

2.5 Extension of existing statistical methodology – aims of this thesis

2.5.1 Approach – spatial point process modelling for plant communities

As indicated above, we will use spatial point process methodology to analyse and model the pattern formed by a plant community trying to infer information on the processes that caused this pattern from the pattern itself.

We are aware that an increasing number of data sets similar to the study data set and even more complex ones will become available (see for example Burslem et al. (2001)) and that sufficient methodology has not yet been developed to analyse these. Furthermore, as established in Sections 2.1.6 and 2.3.2, we believe that spatial statistical methods provide an interesting and valuable alternative approach to individual-based models, which have recently been applied in ecology, by addressing a number of their shortcomings.

The main challenge that we face when dealing with the current data set is its complexity, i.e. the high number of different species. Methodology has to be reconsidered keeping this issue in mind. To illustrate this point, Figure 2.6 shows a plot of the locations of all individuals in the data set where different colours represent different species. It is virtually impossible to detect any spatial structure by visual inspection of this plot. Hence, even the graphical display of the data and its interpretation has to be reconsidered. The further analysis becomes similarly difficult, as parameter estimation and simulation of complex spatial point process models is computationally expensive.

Highly complex models may render estimation infeasible due to extremely long running-times (Møller and Waagepetersen 2003b).

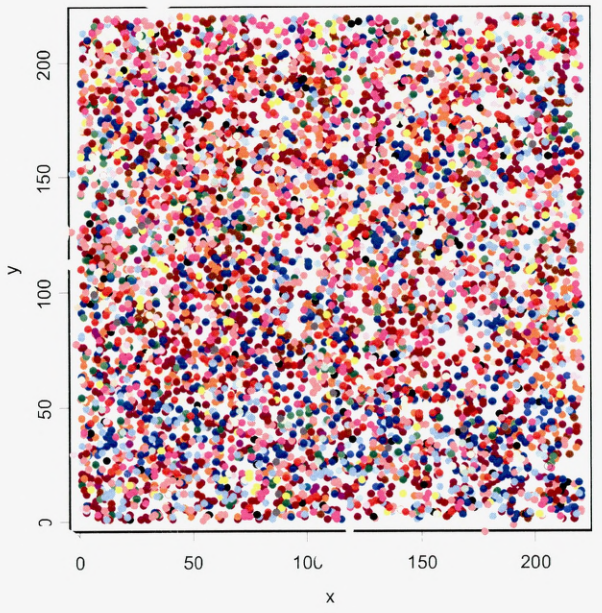

Figure 2.6: Plants by colour

As a result, the aims of this thesis are three-fold. We aim to:

1. Assess the suitability of spatial point process methodology in the context of plant communities;

2. draw conclusions from the results for an understanding on species coexistence and biodiversity in general and;

3. detail concrete results for the specific data set.

2.5.2 Suitability of methodology

2.5.2.1 Exploratory data analysis

We commence analysis of the spatial pattern with an exploratory data analysis to understand the general spatial behaviour of the species and the interactions among them. In view of the high number of species, this is already a very complex task and we assess the suitability of an automated filtering method that can tell "interesting", i.e. non-random, patterns from less interesting patterns by simplifying using methods from statistical quality control (see Chapter 4).

In the same chapter we also address the issue of efficiently assessing the spatial homogeneity of the pattern by refining statistical tests for homogeneity.

2.5.2.2 Reduction of complexity

Clearly, a statistical method able to reduce the dimensionality of the dataset by grouping the point processes according to their spatial behaviour is needed here. However, even though multivariate statistical methods such as principal component analysis (PCA) are well-established methods for dimension reduction in traditional statistics no similar method exists for spatial point pattern data. PCA would yield an overall characterisation of the spatial behaviour of all species in the data set by revealing those aspects of spatial behaviour that the species differ in most. To this end, we develop PCA methodology based on functional principal component analysis and characterise

the community by analysing the shape of those principal components that explain a large proportion of the variance (see Chapter 5).

2.5.2.3 Modelling approach

In theory, it would be possible to fit a model that incorporates all possible intra- and inter-species interactions as well as interactions between the plants and the environment to a data set, given that the necessary information is contained in the data. However, parameter estimation utilises highly complex estimation procedures and algorithms that tend to run for longer than is realistically feasible. Furthermore, the results of such an algorithm would be far too complicated to interpret. In this thesis, we aim to apply a parsimonious modelling approach based on biological information such that parameter estimation is feasible and the results are interpretable. Chapter 6 summarises two modelling attempts and the results obtained.

2.5.3 Biodiversity

As established in Section 2.3.2 we highlight the contributions to ecological theory on biodiversity using spatial point process analysis on the data set described in Section 2.4.

2.5.3.1 Exploratory data analysis

In the initial exploratory analysis we reveal any differences in spatial behaviour between the species. This might give an indication as to whether species have developed niche behaviour. This is particularly interesting as we are dealing with an ancient community. Furthermore, a tendency to ag-

gregation rather than randomness in spatial location would support the niche theory, indicating that conspecifics tend to be closer to each other.

2.5.3.2 Reduction of complexity

In Chapter 5, we aim to reveal those aspects of spatial behaviour that vary most among species. Through this, we are able to reveal that spatial behaviour varies between species, which indicates a niche development. We are also able to describe what aspects of the spatial behaviour yield the best differentiation among species and have thus led to the strongest speciation with regard to spatial behaviour.

2.5.3.3 Modelling approach

In the modelling approach we estimate the strength and the direction (i.e. positive or negative) of the inter-species interaction. We reveal that these vary across the species, providing evidence in favour of the niche theory.

2.5.4 Specific data set

All the above points are also interesting for the specific data set in particular with regard to the re-establishment and conservation of the community.

In addition, we provide methods that may also be successfully applied to other similar communities and be generalised to be applicable to other systems.

2.5.4.1 Exploratory data analysis

The analysis of the subpatterns detects those species that form a clustered, random or regular pattern, as well as determining whether the distribution

of the individuals is homogeneous across the plot or exhibits a trend or another form of inhomogeneity. This will provide valuable information for the rehabilitation process of the specific sample assessed here, but also for other locations in the area where mining is taking place. The knowledge of a species' typical spatial behaviour enables the replanting in a fashion that is as similar as possible to the natural setting. This is likely to facilitate rehabilitation by increasing the survival of the individuals, in particular in an area where the soil is similar and as impoverished as here and species survival is delicately linked to environmental conditions.

2.5.4.2 Reduction of complexity

As noted, the reduction of the complexity of the data set and hence identifying those aspects of spatial behaviour that vary most among the species provides information valuable for the rehabilitation process. These seem to be most critical for community dynamics and thus must be taken into account in the re-planting process.

2.5.4.3 Modelling approach

The information on the strength and direction of inter-species interaction further informs rehabilitation. For instance, species with negative interactions should not be planted in close proximity as competition will weaken the individuals and eventually render them non-viable. On the other hand, species with positive interaction, i.e. those that attract each other should be planted in close proximity to each other.

Chapter 3 will yield a general introduction to point process modelling. Chapter 4 will introduce a novel parsimonious method for second order exploratory data analysis (CUSUM approach) of multi-species spatial point patterns. Chapter 5 introduces principal component analysis (PCA) for highly multivariate spatial point patterns based on functional data analysis methods to reduce the dimensionality of the data set. Chapter 6 describes a hierarchical point process model that incorporates the asymmetric interaction between different plant species with different post-fire regeneration strategies. At the end of each of the chapters the three aspects of the aims of this thesis will be discussed and Chapter 7 will summarise these discussions and suggest further work, partly in view of the applicability of the developed methods to the data collected in the CTFS plots.

Chapter 3

Introduction to Spatial point processes

The statistical methods used in this thesis adapt existing and develop new techniques in the context of spatial point process theory. As a background for this, we provide here a detailed introduction to spatial point processes, which have been conceptually introduced in Section 2.2.2.

Spatial point processes are stochastic models that describe the spatial pattern formed by the locations of objects in two- or more dimensional space. The locations of the objects are regarded as points and the aim of the model is to describe the pattern formed by these points. In this thesis, spatial point process methods are applied to the point pattern formed by individuals within a plant community. The model parameters describe inter- and intra-species interactions and plant densities.

Note that this chapter will include some rather technical definitions and results, assuming substantial knowledge of general mathematics and statistics, including some basic measure theory and the Bayesian approach to statistical inference. They are meant to define rigorously and justify the ap-

proaches but they are not essential for the overall understanding of the thesis. It is beyond the scope of this work to explain the entire underlying general concepts to an extent that they may be understood by a non-specialist without a profound background in statistics. The non-specialist reader is advised to skip the technical paragraphs and focus on the paragraphs describing the rationale for the technical sections that precede each technical definition or explanation, as these convey the main ideas and are sufficient for the understanding of the methods and results presented. A very rigorous but theoretical introduction to the general theory of point processes may be found in Daley and Vere-Jones (1988). Overviews of the theory of spatial point processes are given in Møller and Waagepetersen (2003a,b); van Lieshout (2000); Stoyan and Stoyan (1994); Stoyan et al. (1995). More accessible introductions to spatial point processes for the non-specialist reader may be found in Stoyan and Penttinen (2000) and Diggle (2003).

Also, note that in this chapter, and throughout this thesis, the term "point" will refer to the position of an individual object (in x- and y-coordinates), which in the application described here is the position of an individual plant. The term "location" refers to any position in the area under investigation which does not necessarily coincide with the (x, y)-position of one of the objects forming the pattern. The notation used in this thesis is broadly similar to that used in Møller and Waagepetersen (2003a,b).

3.1 Definition

The general characteristics of spatial point patterns and the importance of these characteristics for the analysis process are discussed in Section 3.1.1. In Section 3.1.2, the concept of a point process model is introduced, followed by a technical definition of spatial point processes (3.1.3) and a coverage of marked point processes (3.1.4).
In the remainder of this chapter, we introduce methods that have been used to aid the following steps for an analysis:

1. **Describe the first and second order characteristics** of a pattern (see Section 3.3).

2. **Select an appropriate spatial point process model** for the pattern, informed by the results of step 1 (see Sections 3.2 and 3.4).

3. **Fit the selected model** to the pattern by estimating the specific parameters (see Section 3.5).

4. **Assess the goodness of fit** of the model by running simulations and then comparing the simulated patterns with the original pattern, based on the first and second order characteristics of each (see Section 3.6).

3.1.1 Characteristics of spatial point patterns

In general, the spatial point pattern formed by the position of objects can exhibit a number of characteristics. These fall broadly into two classes:

i) **"first order" characteristics, describing the density of points through space**

The density of points may be constant over the area considered, thus resulting in a homogeneous pattern, or instead exhibit spatial trend or other forms of non-constant density, resulting in an inhomogeneous pattern.

ii) **"second order" characteristics, relating to the relative position or interaction among points**

Points may be randomly scattered in space, show regularity resulting from negative interaction (inhibition) among the points, or exhibit a clustered structure (aggregation) resulting from positive interaction (attraction) among points.

Figure 3.1 (a) shows the spatial point pattern formed by the species *Scholtzia involucrata* which appears to have a greater plant density towards the bottom of the plot, i.e. showing inhomogeneity and Figure 3.1 (b) shows the spatial point pattern formed by the species *Dasypogon bromeliifolius* which appears to be clustered. Both these patterns were taken from the Cooljarloo data set introduced in Section 2.4.

The first and second order characteristics of a spatial pattern play a central role, and in fact form the first and crucial step, in the analysis of a spatial point pattern. Section 3.3 details how these characteristics are formally assessed.

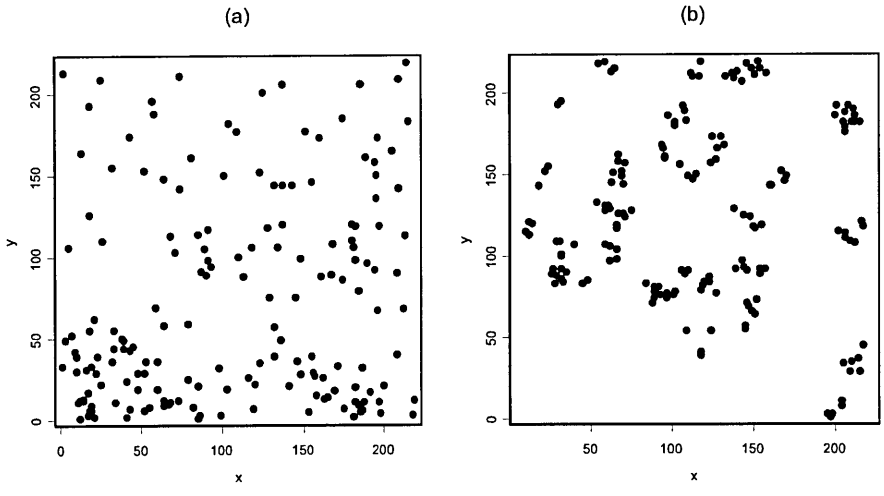

Figure 3.1: Example of (a) an inhomogeneous point pattern, species *Scholtzia involucrata* and a clustered point pattern, species *Dasypogon bromeliifolius* (b).

3.1.2 Point process models

A **spatial point process** is a mathematical model that captures the characteristics of spatial point patterns in a finite number of parameters (Stoyan et al. 1995). A specific model with a given set of parameters provides a mechanism by which **spatial point patterns** may be generated, all having the same spatial characteristics. A spatial pattern generated from a spatial point process is termed a **realisation** of the process. Note that in this thesis we only deal with spatial patterns in the two-dimensional space; the methodology can easily be generalised to patterns in more general spaces and higher

dimensions.

A spatial point process model is stochastic, and so two realisations of the same process, i.e. two patterns generated by the same mechanism, will not be the same but merely share common characteristics with regard to the distribution of the points in space. Figure 3.2 shows examples of two patterns generated from the same process – a homogeneous Poisson process that may be used to model a completely random distribution of points in space (see Section 3.2 for more details on homogeneous Poisson processes).

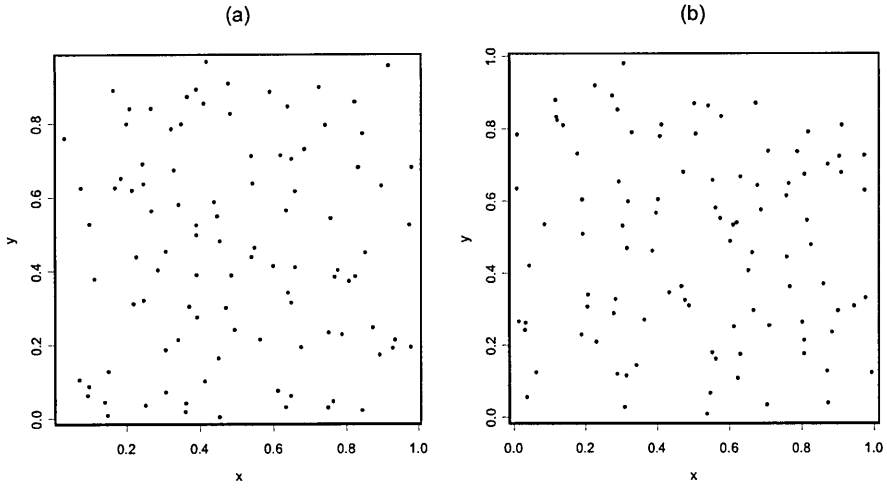

Figure 3.2: Example of two spatial patterns, each generated from the same spatial point process with an expected average number of points 100

More specifically, the mathematical formulation of a spatial point process contains parameters which reflect the characteristics of the process and ultimately the characteristics of the generated patterns, i.e. characteristics such as clustering, regularity or randomness as well as homogeneity or inhomogeneity.

Spatial point processes which differ only by the parameter values in their mathematical formulation are termed **spatial point process models**. These are grouped into **classes** of spatial point process models with similar mathematical descriptions and properties. Various model classes have been formulated in the literature (see e.g. van Lieshout (2000), Møller and Waagepetersen (2003b), Stoyan and Stoyan (1994)); in Sections 3.2 and 3.4 we give a brief overview of a selection of some important models and classes of models relevant here.

Figure 3.3 illustrates three examples of homogeneous patterns generated from different point process models. The pattern in Figure 3.3 (a) has been generated using a homogeneous Poisson process (see Section 3.2), a spatial point process model that describes complete spatial randomness. The pattern in Figure 3.3 (b) has been generated using a Strauss process (see Section 3.4.4), a model of repulsion or spatial regularity (Strauss 1975). The pattern in Figure 3.3 (c) has been generated using a Poisson cluster process (see Section 3.4.4), a spatial point process that may be used to model certain types of clustered patterns (Stoyan et al. 1995).

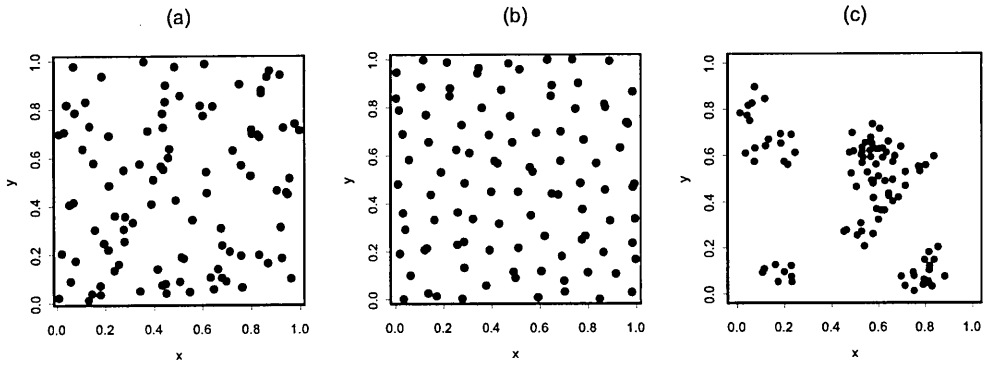

Figure 3.3: Example of a random pattern (a), a regular pattern (b) and a clustered pattern (c)

3.1.3 Technical definition

A spatial point process is a random variable X, with an observed pattern being a realisation x of this random variable. For each Borel set $B \subset \mathbb{R}^2$, let $\phi_X(B)$ be the number of points of X in B[1]. Here, we identify a **point configuration** with a counting measure ϕ_X on Borel sets on $\mathbb{R}^2$.

[1]As mentioned above more general spaces can be considered such as the $\mathbb{R}^d$ or other metric spaces equipped with a metric $d(.,.)$ which are Polish, i.e. complete and separable. For details see Daley and Vere-Jones (1988).

Let N be the set of all such measures. On N define $\mathcal{N}$ as the smallest σ-algebra generated by sets of the form

$$\{\phi \in N : \phi(B) = n \quad n \in \{0, 1, 2, \ldots\}, B \text{ some bounded Borel set}\}.$$

Let $(\Omega, \mathcal{A}, P)$ be some probability space. A **spatial point process** X may then be regarded as a measurable mapping from $(\Omega, \mathcal{A})$ into $(N, \mathcal{N})$, i.e. as a random variable. A spatial point pattern x is then a realisation of this random variable. To avoid unnecessary notation we will not distinguish between X and the measure ϕ_X defined by it, i.e. $X(B)$ denotes the number of points X has in B.

In applications, the process X lives in some subset W of $\mathbb{R}^2$ and patterns are only observed in a bounded area $S \subset W$. Even so, S can be of a very general form. In this thesis S is assumed to be a rectangular set $S \subset W \subset \mathbb{R}^2$ without loss of generality as this is the situation relevant for the study dataset. By this, we also avoid unnecessary complications in the notation in later sections. Individual points in X will typically be denoted by ξ and η. Locations in S which may or may not coincide with a point in X will be denoted by u.

Note that we assume X to be **simple**, i.e. not more than one point may occur in any location.

3.1.4 Marked point patterns

In situations where additional data exist on the objects that form the spatial point pattern under investigation, these additional data are conventionally termed **marks**. Combining a spatial point process with marks yields a

marked pointed process and hence marked point patterns. The marks may be either quantitative variables, such as weight or height in the case of plants, or qualitative variables such as species, thus defining different **types** of points in the pattern.

More formally, let Z be a simple point process in $\mathbb{R}^2$. Attach a random mark $m_\xi \in \mathcal{M}$, where $\mathcal{M}$ is some mark space, to each point $\xi \in Z$. This yields a marked point process

$$X_m = \{(\xi, m_\xi) : \xi \in Z\}.$$

In most applications, the mark space M is a subset of $\mathbb{R}^d$ with $d \geq 1$, but more general mark spaces may be considered, see Schlather (2001) or Stoyan and Stoyan (1994).

Note that if $\mathcal{M} = \{1, \ldots, k\}$, X is a multi-type point process with k different types of points. Note further that a multi-type process can also be regarded as a k-tuple of different subprocesses $(X_1, ..., X_k)$.

The data set analysed in this thesis is a multi-type point pattern consisting of 67 sub-patterns, the positions of 67 different species, i.e. here $k = 67$ and $\mathcal{M} = \{1, \ldots, 67\}$. As noted, the multi-type patterns analysed in the literature comprise points of at most two or three types; the methodology described in Chapters 4, 5 and 6 has been developed in order to deal with the situation of a highly multivariate data set.

3.2 The homogeneous Poisson process – a null model

3.2.1 Overview

We now define a null model, the spatial **homogeneous Poisson process**, which may be used to model patterns that we have previously termed random patterns (see Section 2.1). These patterns are more correctly referred to as patterns exhibiting complete spatial randomness (CSR).

Other more complicated models, including those defined in Section 3.4 below can be viewed as a generalisation of the homogeneous Poisson process which may be used to construct more complicated models. It also serves as a reference process when the first and second order spatial characteristics of a specific pattern are analysed (see Section 3.3 for details).
Generally speaking, a homogeneous Poisson process X has the following two properties:

(1) The density of points is constant in the area under investigation.

(2) The location of any point in the pattern is independent of the location of any of the other points in the pattern, i.e. there is no interaction between the points.

3.2.2 Technical definition

Definition 1 mathematically formalises the properties (1) and (2) in section 3.2.1 above.

Definition 1 *Let μ be the Lebesgue measure on the Borel sets in S. Then the Poisson process on S with intensity λ_0 has the following properties:*

(1) For any Borel set $B \in S$ the cardinality of B, $X(B)$, follows a Poisson distribution with mean $\lambda_0 \cdot \mu(B)$.

(2) For any disjoint Borel sets $B_1, \ldots, B_{n_0} \subseteq S$ with an arbitrary $n_0 \geq 2, X(B_1), \ldots, X(B_{n_0})$ are independent.

The second property is sometimes called the **independent scattering property** (Møller and Waagepetersen 2003b). Note that this property is equivalent to saying that conditional on $X(B) = n$, the n points in X are mutually independent; the spatial point process model with a fixed number of points constructed in this way is called a **binomial point process**.

3.2.3 Simulation of Poisson processes

It is often necessary to simulate a pattern from a spatial point process model, e.g. for the purpose of model validation (see Section 3.6).
Simulating a pattern x from a **homogeneous Poisson process** X with intensity λ_0 is straightforward and consists of two steps (Stoyan and Stoyan 1994):

1. Generate n_{sim}, the number of points in the simulated pattern, by simulating from a Poisson distribution with mean $\mu(S) \cdot \lambda_0$. Standard statistical packages such as R or S-PLUS have a built-in function to do this.

2. Generate two vectors of n uniformly distributed values for the x-coordinates and y-coordinates of the points, respectively.

We shall see in Section 3.4 that simulations become more complicated as the generality of the models increases such that advanced simulation techniques will be necessary.

We shall also see in Chapters 4, 5 and 6 that in the Cooljarloo data set only a few of the subprocesses can be modelled with a simple homogeneous Poisson process, the others require more complicated modelling approaches. In addition, when we are aiming at modelling the whole data set multi-type models will have to be applied, see Sections 3.4.4.3 and 3.4.9 for more details.

3.3 Summary Statistics

When analysing a spatial point pattern, we seek to infer properties of the pattern. These properties will then be used to choose the appropriate point process model. Realisations of this model are patterns with properties similar to the original. Hence, the initial analysis of a spatial point pattern consists of a descriptive analysis of the general properties of the pattern using summary statistics. The information gained from the summary statistics will subsequently allow the choice of a specific point process model to fit to the data.

The summary statistics reflect the characteristics as described in Section 3.1.1; i.e. in analogy to first and second order characteristics, we distinguish between **first and second order summary statistics**[2].

[2]Third and higher order summary statistics have also been considered, see Schladitz and Baddeley (2000), but will be omitted here.

3.3.1 First order summary statistics

First order summary statistics are an analogue to the mean in standard, i.e. non-spatial, statistics in the sense that they describe the density or intensity of the spatial pattern in terms of the average number of points per unit area. However, the average number of points varies in space if the process is inhomogeneous. Hence, unlike in standard statistics, the first order summary statistic is not a single value but has to be regarded as a function of the spatial location, i.e. a function of x and y, and can be plotted as a 3-dimensional surface.

Figure 3.4 shows an example of an inhomogeneous pattern with a strong trend and its estimated intensity surface (for details on intensity estimation see Section 3.3.3).

More technically, for a point process X the **intensity measure** is given by

$$\Lambda(B) = E[\phi(B)], \quad \text{for any Borel set } B,$$

where ϕ is a counting measure as defined above and $E[.]$ denotes the expected value. If Λ is absolutely continuous with regard to the Lebesgue measure, a density function, the **intensity function** $\lambda : B \rightarrow \mathbb{R}^+$ exists, such that

$$\Lambda(B) = \int_B \lambda(x)dx.$$

If the intensity is constant, i.e. if $\lambda(x) = \lambda_0$ the point process X is called a **homogeneous** or **(first order) stationary** process.

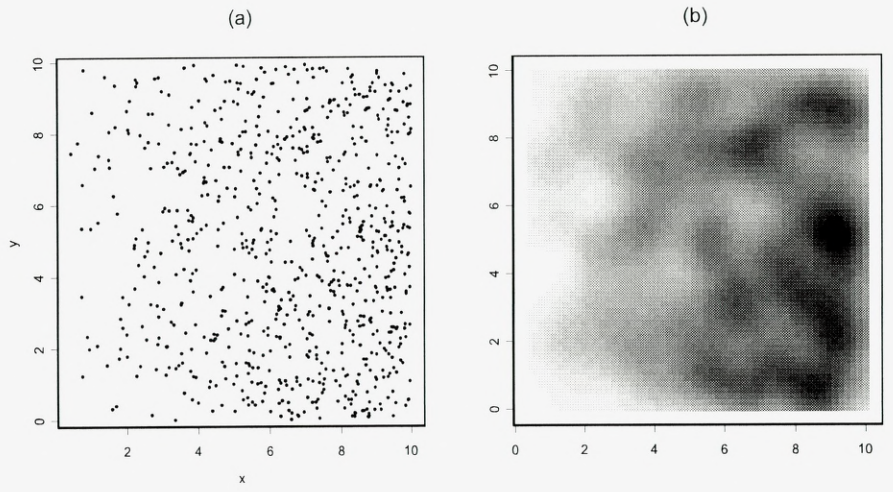

Figure 3.4: Inhomogeneous point pattern (a) and estimated intensity surface for the same inhomogeneous point pattern (b)

3.3.2 Second order summary statistics

Second order summary statistics are an analogue to measures of dispersion in non-spatial statistics, in the sense that they describe the location of an individual point relative to the other points. Since we consider the expected number of points at and within specific distances from each point, this class of summary statistics does not yield a single measure but is a function of distance. A number of second order summary statistics have been suggested in the literature. We will only introduce the most commonly applied statistics. These are Ripley's K-function (Ripley 1976) and its variant, the L-function

(Møller and Waagepetersen 2003b), as well as the pair correlation function (Stoyan et al. 1995)[3].

For a homogeneous process with $\lambda(x) = \lambda_0$ **Ripley's K-function** considers the expected number of points within a distance r from an arbitrary point and is defined as

$$K(r) = \frac{\text{expected no. of points } \textbf{within} \text{ dist. } r \text{ from arbitrary point}}{\lambda_0}. \quad (3.1)$$

Often, the variance stabilising L-function

$$L(r) = \sqrt{\frac{K(r)}{\pi}} \quad (3.2)$$

is used instead (Besag 1976; Møller and Waagepetersen 2003a).
The pair correlation function g considers the expected number of points at a distance r from an arbitrary point.

$$g(r) = \frac{\text{intensity of points } \textbf{at} \text{ dist. } r \text{ from arbitrary point}}{\lambda_0}. \quad (3.3)$$

The pair correlation function and the K- (and similarly the L-) function are clearly related; the K-function may be regarded as a cumulative version of the pair correlation function.

In the inhomogeneous case, the formula (3.1) has to be generalised since the intensity $\lambda(x)$ is no longer constant. This is achieved by considering the local intensity $\lambda(x)$ for each point (see Baddeley et al. (2000)) but the construction principle is still the same.

[3]Other second order summary statistics that have been discussed in the literature include the G- and F function, see Diggle (2003), as well as the J-function, see van Lieshout and Baddeley (1999).

Note that the second order summary statistics may be used to distinguish clustered, random and regular patterns. For a homogeneous Poisson process with complete spatial randomness we have $K(r) = \pi r^2$, $L(r) = r$ and $g(r) = 1$. Thus, if $K(r) > \pi r^2$ at close distances we have a clustered pattern as we find on average more points within these distances than expected from a random pattern. $K(r) < \pi r^2$ indicates a regular pattern. Similarly, if $L(r) > r$ we have a clustered pattern and if $L(r) < r$ a regular pattern. Analogously, the pair-correlation function with $g(r) > 1$ indicates clustering and $g(r) < 1$ regularity.

The theoretical form of both types of second order summary statistics is known only for a small number of special cases of models other than the homogeneous Poisson process. In general applications, estimated functions along with the function for a homogeneous Poisson process are plotted against distance. We discuss the estimation of the summary statistics in Section 3.3.3. Example plots are thus omitted here. Refer to Figure 3.7 in Section 3.3.3 for examples.

3.3.2.1 Homogeneous case

More technically, for a homogeneous process with $\lambda(x) = \lambda_0$ and with finite intensity λ_0 we define **Ripley's K-function** (Ripley 1976) as

$$K(r) = E \sum_{\xi \in X,\, \eta \in X, \xi \neq \eta} \mathbf{1}[\|\xi - \eta\| \leq r]/(\lambda_0^2 \times \mu(S)), \tag{3.4}$$

where $\mathbf{1}(.)$ denotes the indicator function and μ the Lebesgue measure. For two Borel sets B_1 and B_2 we define the **second order factorial moment**

measure $\mu^{(2)}$ as

$$\mu^{(2)}(B_1 \times B_2) = E \sum_{\xi \in X, \eta \in X, \xi \neq \eta} \mathbf{1}[\xi \in B_1, \eta \in B_2].$$

If $\mu^{(2)}$ has a density function $\rho(2)$, i.e. if

$$\mu^{(2)}(B_1 \times B_2) = \int_{B_1} \int_{B_2} \rho^{(2)}(\xi, \eta) d\xi \, d\eta$$

then $\rho(2)$ is called the second-order product density. If, in addition, X is stationary and isotrophic such that $\rho^{(2)}(\xi, \eta) = \rho(r)$, where $r = \|\xi - \eta\|$, the **homogeneous pair correlation function** is defined as:

$$g(r)_{\text{hom}} = \frac{\rho(r)}{\lambda^2}.$$

See Stoyan and Stoyan (1994); Stoyan et al. (1995); Møller and Waagepetersen (2003b) for more details on both the K-function and the pair-correlation function[4].

3.3.2.2 Inhomogeneous case

Baddeley et al. (2000) introduce an inhomogeneous version K_{inhom} of the K-function and hence an inhomogeneous version L_{inhom} of the L-function by taking local intensities into account[5]. Here $\lambda^2(S)$ is replaced by $\lambda(\xi)\lambda(\zeta)$ to yield

$$K_{\text{inhom}}(r) = E \sum_{\xi \in X, \, \eta \in X, \xi \neq \eta} \mathbf{1}[\|\xi - \zeta\| \leq r]/(\lambda(\xi)\lambda(\zeta)). \tag{3.5}$$

[4]Note that the homogeneous pair correlation function is only a special case of the pair correlation function defined below. The special case has typically not been applied in the literature and is thus not considered further in this thesis.

[5]A modification of the K-function has been considered in Stoyan and Stoyan (1994) to account for anisotropy, but will not be considered here.

Again, the variance stabilising L-function tends to be used, where

$$L_{\text{inhom}}(r) = \sqrt{\frac{K_{\text{inhom}}(r)}{\pi}}.$$

The pair correlation function is defined as

$$g(\xi, \eta) = \frac{\rho^{(2)}(\xi, \eta)}{\lambda(\xi)\lambda(\eta)}.$$

Note that the first and second order summary statistics are not unique; patterns derived from two different point process models can have the same first or second order summary statistic (Baddeley and Silverman 1984; Baddeley et al. 2000). Further note that the second order summary statistics are invariant under independent thinning, i.e. their values do not change when points of the process are randomly deleted, independent of their location (Møller and Waagepetersen 2003a). This will be relevant in a simulation study in Chapter 5.

3.3.3 Estimation of summary statistics

For an observed pattern, the true values of the first and second order summary statistics are unknown and have to be estimated from the data. Due to the constant intensity, estimation procedures are simplified when homogeneity can be assumed so we treat homogeneous and inhomogeneous patterns separately.

3.3.3.1 Homogeneous patterns

The intensity can be estimated by simply dividing the number of points in the pattern by the area of S, i.e.

$$\hat{\lambda}_0 = n(x)/|S|. \tag{3.6}$$

This intuitive estimator is in fact the maximum likelihood estimator if X is a homogeneous Poisson process.

Edge effects have to be taken into account, since we only observe the pattern x on a bounded window but points may interact with other points outside this window.

The K-function is typically estimated using an estimator introduced by Ripley (Ripley 1976):

$$\hat{K}(r) = n(x)^{-2}|S| \sum_{\xi,\eta \in x}^{\neq} w_{\xi,\eta}^{-1} \mathbf{1}[\|\xi,\eta\| \leq r],$$

where n is the number of points in region S with area $|S|$, and $w_{\xi,\eta}$ an edge correction factor – the proportion of the circle with centre ξ passing through η which lies in S. This estimator is unbiased for homogeneous processes (Møller and Waagepetersen 2003b).

3.3.3.2 Inhomogeneous patterns

The situation becomes more complicated if inhomogeneous patterns are considered as mentioned above. The intensity at a given location is estimated by taking the number of points within a certain distance into account. In the inhomogeneous case, the intensity is typically estimated using nonparametric kernel estimates, i.e.

$$\hat{\lambda}_b(u) = \sum_{\eta \in x} k_b(u-\eta)/c_{S,b}(\eta).$$

Here, k_b is a kernel with bandwidth $b > 0$ and

$$c_{S,b}(\eta) = \int_S k_b(u - \eta)du$$

is an edge correction factor.

In this thesis, a product kernel will be used, with $k(\xi) = e(\xi_1)e(\xi_2)$ for $\xi = (\xi_1, \xi_2) \in \mathbb{R}^2$, where

$$e(u) = (3/4)(1 - |u|)\mathbf{1}[|u| \leq 1], \qquad u \in \mathbb{R},$$

is the Epanečnikov kernel (Stoyan et al. 1995). For other choices of kernel functions, see Stoyan and Stoyan (1994). Note that the choice of bandwidth is crucial whereas the choice of the kernel function is not; different bandwidths lead to very different estimated intensity surfaces which is not the case when different kernel functions are chosen (Møller and Waagepetersen 2003b). Figure 3.5 shows the estimated intensity surface for the pattern formed by the locations of the species *Scholtzia involucrata* in the Cooljarloo data set collected on a 22m by 22 m plot. In Figure 3.5 (a) a bandwidth of 20 cm, in (b) a bandwidth of 60 cm and in (c) a bandwidth of 100 cm has been used. The plots illustrate that larger bandwidths lead to smoother surfaces but ignore local features. In applications, it is crucial to find the right balance between smoothing out local details and revealing the overall structure. One typically intends to assess the overall intensity of a pattern. Hence, the bandwidth should reflect seed dispersal radii such that local clustering resulting from seed dispersal does not appear in the estimated intensity surface. For the Cooljarloo data set most patterns proved to be inhomogeneous (see chapter 4) so a kernel estimate had to be used. The estimated surfaces for

each species were carefully assessed and in most cases a bandwidth of 60 cm was chosen as it still reveals some local features but also the overall trend.

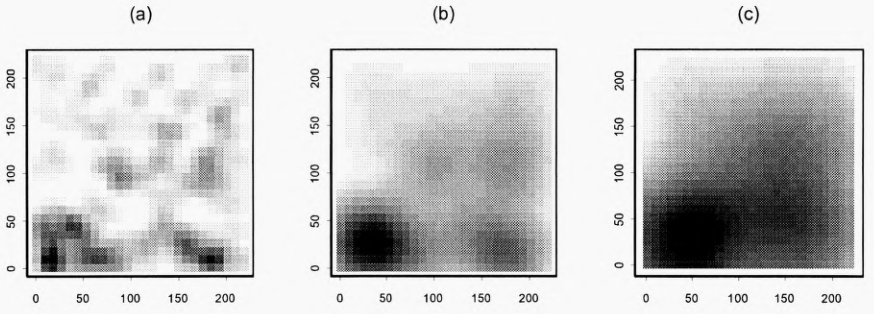

Figure 3.5: Estimated intensities of inhomogeneous point pattern formed by the species *Scholtzia involucrata* using different bandwidths, measurement unit is 10 cm.

Note that edge effects can occur when intensity estimation is done using a kernel function. These are accounted for by including edge correction factors. Figure 3.6 shows the estimated intensity surface for the species *Astroloma xerophyllum* without (a) and with (b) edge correction. It is clear that in Figure 3.6 (a) the intensity appears to get lower closer to the edges, this is not

the case in Figure 3.6 (b). Nevertheless, the intensity might be overestimated in some parts of the plot, consider for instance the bottom left hand corner of the plot.

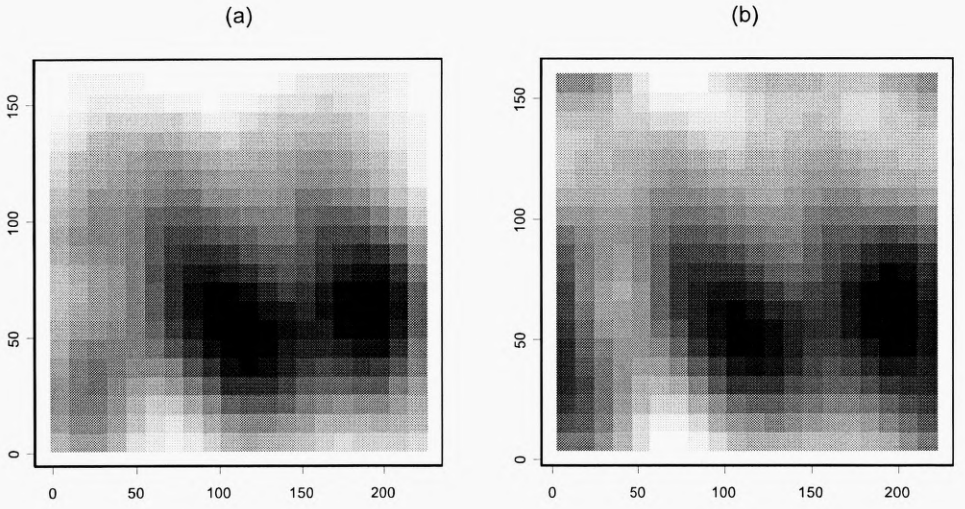

Figure 3.6: Estimated intensities of inhomogeneous point pattern formed by species *Astroloma xerophyllum* without (a) and with (b) edge correction, measurement unit is 10cm.

For the K-function, we use the estimator introduced in Baddeley et al. (2000):

$$\hat{K}_{\text{inhom}}(r) = 2 \sum_{\xi,\zeta \in x}^{\neq} \mathbf{1}[\|\xi - \zeta\| < r]/(|S_\xi \cap S_\zeta| \bar{\lambda}(\xi)\bar{\lambda}(\zeta)),$$

where $\bar{\lambda}$ is the modified intensity estimator

$$\bar{\lambda}_b(u) = \sum_{\eta \in x \setminus u} k_b(u - \eta)/c_{S,b}(\eta),$$

where $\eta \in x \setminus u$ indicates that the sum is over all η in x except u and $C_{S,b}(\eta) = \int_S \kappa(u - \eta)du$ is an edge correction factor and κ a kernel function. This modification is discussed in more detail in Baddeley et al. (2000). Similarly, we use

$$\hat{g}(r) = 2 \sum_{\xi,\zeta \in x}^{\neq} \tilde{k}(r - \|\xi - \eta\|) / \left(\pi r |S_\xi \cap S_\eta| \widehat{\lambda(\xi)\lambda(\eta)}\right)$$

for the estimation of the pair-correlation function, where $\widehat{\lambda(\xi)\lambda(\eta)}$ is an estimator of $\lambda(\xi)\lambda(\eta)$, see Baddeley et al. (2000), and $\tilde{k}$ a kernel estimator.

Note that it is difficult to distinguish between aggregation resulting from inhomogeneity, i.e. a first order property and aggregation due to clustering, a second order property, see Diggle (2003). For instance, a perceived inhomogeneity within a given observation window may merely be the result of larger scale clustering. In other words, the above approach seeks to estimate first and second order properties at the same time by taking the local intensity into account while estimating second order properties. Strictly speaking, this cannot be done as these cannot clearly be distinguished.

Also note that in order to facilitate interpretation we often plot the distance r against the estimated L-function $\hat{L}(r) - r$; for the estimation of second order summary statistics see Section 3.3.3. As an example, Figure 3.7 shows the estimated pair correlation function for species *Scholzia involucrata*, which indicates clustering at smaller distances since the estimated function is above one.

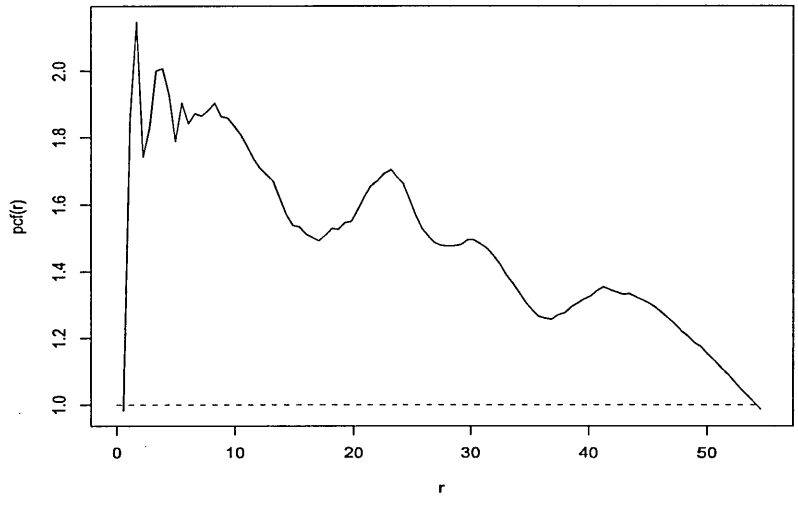

Figure 3.7: Estimated pair-correlation function for species *Scholzia involucrata*, where r denotes distance in units of 10cm.

3.3.4 Summary statistics for multi-type processes

The summary statistics described in Sections 3.3.1 and 3.3.2 can be generalised to a multi-type setting, i.e. to a marked point process with mark set $M = \{1, \ldots, k\}$. In analogy to the univariate case, for two subprocesses X_i and X_j, where $i \neq j, i, j \in M$ the **cross summary statistics** K_{ij}, L_{ij} and g_{ij} consider the number of pairs of points from the two different subprocesses at or within a distance r from an arbitrary point. More formally, we define the multi-type K_{ij}-function for pairs of patterns as

$$K_{ij}(r) = E \sum_{\xi \in X_i,\, \eta \in X_j} \mathbf{1}[\|\xi - \eta\| \leq r]/(\lambda_i(\xi)\lambda_j(\eta)). \tag{3.7}$$

In analogy to the one-dimensional case we may define L_{ij} and g_{ij}. The interpretation of the multivariate second order summary statistics are similar to the interpretation of the univariate second order summary statistics. Note, that it is not possible to determine from these summary statistics whether observed attraction or observed repulsion between points is due to mutual attraction (or repulsion) or due to asymmetric interaction.

Also, these summary statistics may be generalised to more than two types of points, but interpretation becomes increasingly difficult.

3.3.5 Tests of homogeneity and complete spatial randomness

Tests for homogeneity In many applications it is useful to know whether a given spatial point pattern x is consistent with a pattern generated from a homogeneous process. Often, the careful visual inspection of the plotted pattern and the estimated intensity surface may indicate the homogeneity of the pattern. For a more formal test introduced by Ripley (1976), the observation window S is, *a priori*, split into two disjoint subwindows S_1 and S_2. We aim to test the null hypothesis

$\mathbf{H_0}$: The pattern x is a realisation of a stationary process.

versus

$\mathbf{H_1}$: The pattern x is not a realisation of a stationary process.

Ripley suggests the following test statistic:

$$F = \frac{|S_1|(2n_2 + 1)}{|S_2|(2n_1 + 1)}, \tag{3.8}$$

where n_1 and n_2 are the number of points in each of the two subregions and S_1 and S_2 their areas, respectively. F is approximately distributed as

$F_{(2n_1+1,2n_2+1)}$ if x is a realisation of a homogeneous Poisson process (Sachs 1984; Stoyan and Stoyan 1994). Note that this test heavily relies on the choice of the subwindows. In Chapter 4, we suggest an approach based on a permutation test which circumvents this problem. We also consider the application of a Kolmogorov Smirnov goodness of fit test, which yields a more refined approach with finer sub-divisions.

Tests for complete spatial randomness A number of statistical tests have been suggested in the literature which may be used to test formally for complete spatial randomness. We will only describe tests based on the second order summary statistics in this thesis as these will be further explored and generalised in Chapter 4. See Stoyan et al. (1995) for a detailed overview of other approaches.

A test based on the estimated L-function, $\hat{L}$, uses the following test statistic:

$$T = \max |\hat{L}(r) - r|.$$

In this thesis, this test will be referred to as **max-dist** test. Tables of approximate critical values have been derived by simulation (Ripley 1988; Koen 1991). An alternative to using these is the application of a permutation test, where l binomial processes with the same number of points as the pattern under investigation are simulated and the test statistic T is calculated for each of these. The resulting values $T_1, \ldots, T_l$ are then ordered in ascending order. If the value of T calculated for the original data is larger then the value at position $0.95 \times l$ in this order, then the null hypothesis is rejected. Note that this test has been constructed to be used in the homogeneous case.

When applied to an inhomogeneous Poisson pattern the test may indicate significant but spurious clustering. This is due to the fact that in any inhomogeneous pattern there are more points at closer distances in some region of the observation window than in other areas. In Chapter 4, we suggest an alternative approach which can be used to detect quickly patterns that do not exhibit complete spatial randomness where we will expand in more detail on the issue of homogeneous and inhomogeneous patterns in this context. This approach is particularly useful in the context of multi-type point patterns like the Cooljarloo data set.

3.4 Other spatial point process models

The homogeneous Poisson process, as introduced in Section 3.2, is a very simple spatial point process model, and most other models may be regarded as a generalisation of it. It thus serves as a reference model for more complex models. A large number of different models have been introduced in the literature. This section will describe those classes of point process models which will be relevant for the modelling of the Cooljarloo data set, as described in Chapter 6. See for example van Lieshout (2000); Stoyan et al. (1995); Møller and Waagepetersen (2003b) for a more exhaustive overview of other spatial point process models.

3.4.1 Inhomogeneous Poisson

A straightforward generalisation of the homogeneous Poisson process may be achieved through introducing inhomogeneity, but no interaction. Thus the

two properties considered in 3.2.1 now become:

(1) The intensity of the point pattern is **not** constant over the bounded region.

(2) There is no interaction between the points.

The intensity of the pattern is now described by a non-constant intensity function $\lambda(r)$. See Figure 3.8 for a realisation of an inhomogeneous Poisson process with intensity function $\lambda(x, y) = 4x + 2y$. Figure 3.8 (a) shows the pattern and Figure 3.8 (b) the estimated intensity surface for this pattern.

Formally, Definition 2 generalises Definition 1 as follows:

Definition 2 *Let ρ be a locally finite and diffuse measure defined on the Borel sets in S, i.e. $\rho(B) < \infty$ for all Borel sets B on S and ρ has no mass at any point in S. Then the Poisson process on S with intensity measure ρ written as*

$$X \sim Poisson(S, \rho)$$

has the following properties:

(1) For any Borel set B in S the cardinality of B, $X(B)$, follows a Poisson distribution with mean $\rho(B)$.

(2) For any disjoint Borel sets $B_1, \ldots, B_{n_0} \subseteq S$ with an arbitrary $n_0 \geq 2, X(B_1), \ldots X(B_{n_0})$ are independent.

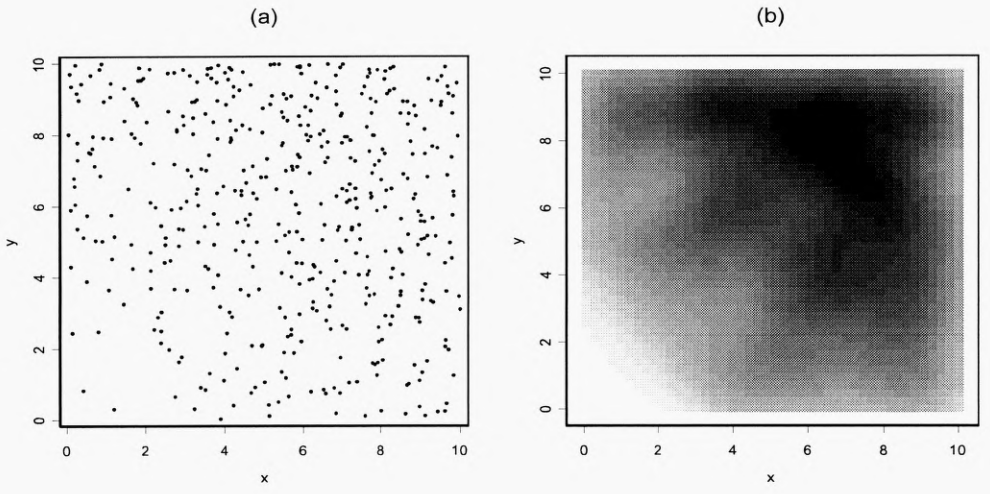

Figure 3.8: Point pattern generated from an inhomogeneous Poisson process with intensity function $\lambda(x, y) = 4x + 2y$ (a) and estimated intensity surface for the same pattern (b).

Note that using this notation, the homogeneous Poisson process with intensity λ_0 may be written as:

$$X \sim Poisson(S, \mu \cdot \lambda_0),$$

where μ is the Lebesgue measure.

If the measure ρ is given by a density λ then we call λ the **intensity function** of the process and write $X \sim Poisson(S, \lambda)$ instead of $X \sim Poisson(S, \rho)$.

Note further that it is possible to characterise a Poisson point process by its **void probabilities** for all Borel sets B, i.e. the following probabilities

(Daley and Vere-Jones 1988):

$$P(X(B) = 0) = \exp(-\mu(B)).$$

The class of Poisson processes is closed under superpositioning and independent thinning (Møller and Waagepetersen 2003a). This property will become relevant when we assess the performance of the methods introduced in Chapter 5 in the context of noisy data.

3.4.2 Densities with respect to the homogeneous Poisson process

It is often convenient to express a more general point process model in terms of its density, notably for the class of Markov processes, see Section 3.4.4 below. This is done by defining these densities with regard to the unit rate homogeneous Poisson process in accordance with the Radon-Nikodym theorem (Møller and Waagepetersen 2003b). The density describes the probability for a point pattern relative to the homogeneous Poisson process, and is not to be confused with the entirely different concept of an intensity, where the expected number of points in a given location is being considered.

Note, that general Poisson processes are always absolutely continuous with respect to the **standard or unit rate Poisson process**, i.e. the Poisson process with constant intensity $\lambda_0 = 1$, when defined on a bounded subset S of $\mathbb{R}^d$ (Møller and Waagepetersen 2003b).

For a general Poisson process

$$f(x) = \prod_{i=1}^{n(x)} \lambda(x_i) \exp\left(- \int_S [\lambda(\mu) - 1] d\mu \right)$$

is the density function (Møller and Waagepetersen 2003b). For the **homogeneous Poisson process** on S with intensity $\lambda_0 > 0$ this simplifies to

$$f(x) = \exp(-(\lambda_0 - 1)|S|) \cdot \lambda_0^{n(x)},$$

with $n(x)$ the number of points in x and $|S|$ the volume of S.
Note that we can use the density representation of the Poisson process to show that, for example, the estimator of the intensity in equation 3.6 in Section 3.3.3.1, is the maximum likelihood estimator of the intensity for a homogeneous Poisson process (Møller and Waagepetersen 2003b). Unlike the densities of all other models in this section the density of the inhomogeneous Poisson process is known in its full form. However, for more complicated models, the density is known only up to a normalising constant.

3.4.3 Simulating inhomogeneous Poisson patterns

The simulation of a pattern x from $X \sim Poisson(S, \lambda)$, an inhomogeneous Poisson process, may be done by thinning from a homogeneous process. Here, we assume that the intensity function of X is bounded, such that we can find a value λ^* such that $\lambda(u) \leq \lambda^*$ for all $u \in S$.

1. Generate a homogeneous Poisson process with intensity λ^*.
2. Delete each of the generated points with probability $p(u) = \frac{\lambda(u)}{\lambda^*}$.

This procedure is sometimes called **location dependent thinning** (Møller and Waagepetersen 2003b).

3.4.4 Markov point processes

The class of spatial Markov point processes[6] models patterns exhibiting aggregation (or inhibition) due to interaction between points (van Lieshout 2000). Referring back to Section 3.2.1 this implies a further generalisation of the properties (1) and (2) in Section 3.2 in as far as that now interaction between points is allowed.

Note that, for brevity, we will use the term Markov process in this thesis instead of spatial Markov point processes. Some authors use the term Gibbs process instead of Markov process, see Stoyan et al. (1995).

3.4.4.1 Strauss process

The simplest example of a Markov process is the **Strauss process** (Strauss 1975). Here, interaction is constant within a fixed interaction radius R around each point. The strength of the interaction can range from no interaction to complete inhibition within the radius R around each point. Note that in the case of no interaction the process is equivalent to a homogeneous Poisson process.

Formally, the density of the Strauss process is given by

$$f(x) = \alpha\beta^{n(x)}\gamma^{s_R(x)}, \tag{3.9}$$

where $n(x)$ is the number of points in X, $s_R(x)$ the number of distinct pairs of points in X with a distance $r < R$ and α the normalising constant. The parameter $\beta > 0$ reflects the intensity of the process and the parameter γ

[6]This class of processes is termed *Markov* due to a spatial Markov property referring to interaction with spatial neighbours only (van Lieshout 2000).

the strength of the interactions between points with a distance $r \leq R$. Note that for $\gamma = 1$, $X \sim Poisson(S, \beta)$.

See Figure 3.9 for realisations of two different Strauss process models with $\beta = 50$. Figure 3.9 (a) is a realisation with $\gamma = 0.95$ resulting in a pattern very close to CSR and Figure 3.9 (b) is a realisation with $\gamma = 0.001$ resulting in a pattern with strong repulsion.[7]

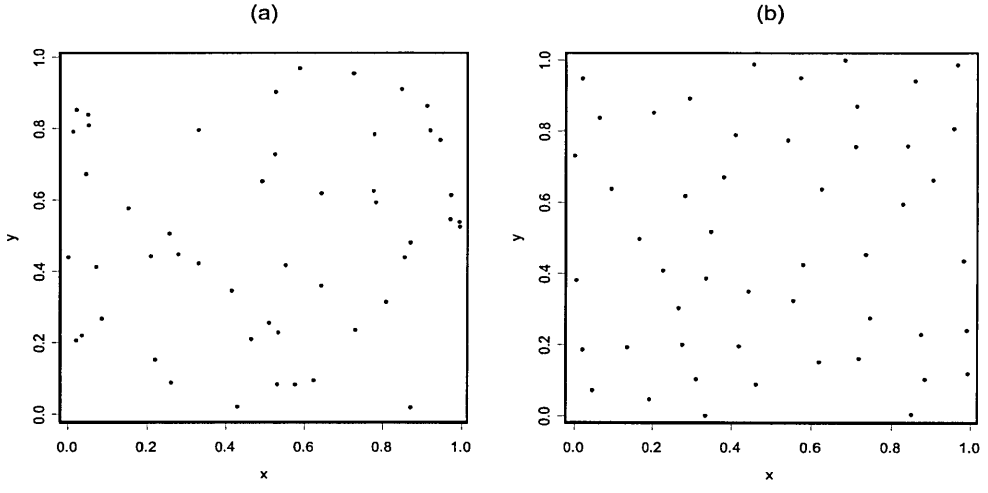

Figure 3.9: Patterns generated from a Strauss process model with different strengths of interaction.

[7]The special case where $\gamma = 0$ is called a **hard core process**. In this case $f(x) = \alpha\beta^{n(x)}0^{s_R(x)} = 0$ if two points have a distance $r \leq R$, i.e. this occurs with zero probability. Otherwise, $f(x) = \alpha\beta^{n(x)}0^0 = \alpha\beta^{n(x)}$.

3.4.4.2 Pairwise interaction processes

A more general class of Markov point processes are the **pairwise interaction processes**. Here the interaction around each point is not constant but changes with the distance from the point according to an interaction function. Its density is of the following form:

$$f(x) = \alpha \prod_{i=1}^{n(x)} \beta(\xi) \prod_{\xi,\eta \in X}^{\neq} h(\|\xi - \eta\|), \tag{3.10}$$

where α is the normalising constant as in equation 3.9, $\beta(.)$ is a function describing the intensity of the process and $h(.)$ is a non-negative interaction function.

A number of interaction functions have been considered in the literature; van Lieshout (2000) gives a detailed overview of various choices of interaction functions. For restrictions on the choice of interaction functions see, for example Diggle (2003). The models suggested in Chapter 6 use interaction functions where the strength of interaction decreases with distance from each point within a specific interaction radius.

3.4.4.3 Multivariate Markov point processes

The univariate Markov point process may be generalised to the bivariate or even multivariate case in a straightforward way. Interaction now takes place between k different types of points. Formally, densities are defined with regard to k independent standard homogeneous Poisson processes. At this stage we will only consider bivariate processes with subprocesses X and Y, which can be easily extended to more than two types of points.

For instance, the **Widom-Rowlinson penetrable spheres mixture model** (Widom and Rowlinson 1970) is defined by the density

$$f(x,y) = \alpha\beta_1^{n(x)}\beta_2^{n(y)}\mathbf{1}\{d(x,y) > R\}.$$

Here, β_1 and β_2 are parameters describing the densities of the two subprocesses and $d(x,y)$ is the minimum of all Euclidean distances between all points in X and all points in Y.

Another class of bivariate spatial Markov point processes is the **Continuum Ising model** (Georgii and Haggström 1996). Here, we consider two processes X and Y. The model is defined by the density:

$$f(x,y) = \alpha\beta_1^{n(x)}\beta_2^{n(y)}\prod_{\xi\in X}\prod_{\eta\in Y}h(\xi,\eta),$$

where h is an interaction function describing the interaction between points from process X and Y and $0 \leq h(.,.) \leq 1$; $n(x)$ and α are defined as in equation (3.10). Note that the marginal distributions for X, Y and $X \cup Y$ are usually complicated but that X conditional on Y (and vice versa) is a Poisson process. This model can easily be extended to more than two processes and is fitted to the Cooljarloo data set in an initial modelling attempt in Chapter 6. We shall see there that the Continuum Ising model is too simple in this situation, so a more complex model has to be used.

See van Lieshout (2000) and Baddeley and Møller (1989) for other examples of marked Markov point process models.

3.4.5 Simulating Markov processes

As mentioned above, the density of Markov processes is typically only known up to a normalising constant. Hence, the simulation of these processes is

usually not straightforward. The most common approach to overcome this problem is to apply Markov Chain Monte Carlo methods, typically based on the **Metropolis Hastings algorithm**, in particular birth and death Metropolis Hastings algorithms, as introduced in Geyer and Møller (1994). Møller and Waagepetersen (2003a,b) describe the various methods in great detail, including more advanced simulation methods such as perfect simulation (Berthelsen and Møller 2002).

3.4.6 Neyman-Scott processes

Neyman-Scott processes, often also referred to as Poisson cluster processes, are a class of spatial point processes termed "mother-daughter" processes by some authors (Diggle 2003; Stoyan et al. 1995). These processes are constructed in two steps.

1. "Mother" points are generated from a Poisson process with constant or non-constant intensity function λ.

2. For each mother, a random number of "daughters" is generated, where the number of offspring is independently identically realised from a univariate distribution for each parent, and the locations of the daughters follow a bivariate distribution around the mother points.

The locations of the points in the daughter process form the Neyman-Scott process. Examples of Neyman-Scott processes include the **Matérn cluster process**, where the locations of the daughter points are uniformly distributed on the area of a circle with radius r_T around the mother points. Similarly, the **Thomas processes** is a point process model, where the locations of

the daughter points follow a bivariate normal distribution with mean 0 and variance-covariance matrix $\sigma_M^2\mathbf{1}$ around the mother plants, where $\mathbf{1}$ denotes the 2×2 unit matrix (Møller and Waagepetersen 2003b).

3.4.7 Simulating Neyman-Scott processes

The simulation of Neyman Scott point processes simply follows the two-step procedure described above. Initially, a homogeneous or inhomogeneous Poisson process with intensity λ is generated as described in Sections 3.2.3 and 3.4.3, respectively, yielding the mother process. As a next step a cluster is then generated around each mother point. This algorithm will be applied repeatedly in a simulation study in Chapter 5 where we assess the feasibility and performance of the methods introduced there.

Figure 3.10 shows two realisations of Matérn processes with constant intensity of the mother process of $\lambda(u) = \lambda_0 = 5$ (a) and $\lambda(u) = \lambda_0 = 20$ (b), respectively, and $r_T = 0.1$ on the unit square.

Neyman-Scott processes may be suitable as a model when the positions of plants from a single species were modelled. One could, e.g. model the location of a seeder or a resprouter species in the Cooljarloo data set. The mothers of the seeder species, i.e. the original plants that shed their seeds but die in the fire, would no longer be visible. When resprouting species (with rhizomatous behaviour) are considered, the mothers may still exist but it may be difficult to distinguish between mothers and daughters. In Chapter 6 we aim at analysing inter-species interaction to gain an understanding of

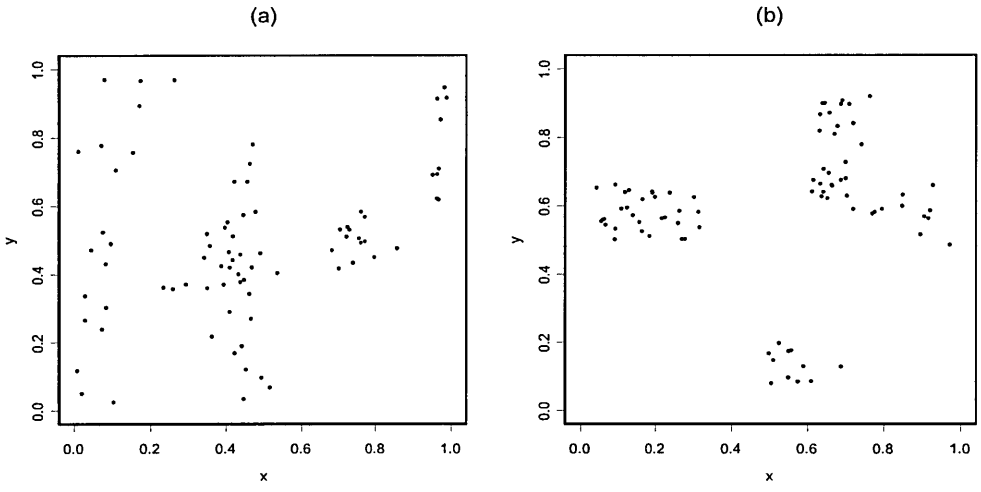

Figure 3.10: Patterns generated from a Matern process with different intensities for the underlying mother process

coexistence so do not apply this type of processes here. However, they are applied in simulation studies in Chapter 5.

3.4.8 Cox processes

In Markov processes, aggregation or inhibition was interpreted as the result of an interaction among the objects represented by the points. **Cox processes** are a class of models describing aggregation or clustering resulting from unobserved environmental variability influencing the location of points in a process. This variability is assumed to be a stochastic process in itself

leading to the Cox processes being called **"doubly-stochastic" processes**, a term introduced in Cox (1955).

For instance, varying but unobserved nutrient levels in the soil may be considered as unobserved environmental variability and thus as a realisation of a stochastic process. Similarly, another spatial point pattern may be considerd as a realisation of a random process influencing the process under investigation. Cox processes are constructed such that the resulting process is a Poisson process, given the underlying random process.
Formally, this yields the following definition:

Definition 3 *Let $Z = \{Z(u) : u \in S\}$ be a non-negative locally finite random field on S. If the conditional distribution of a spatial point process X given Z, denoted by $X|Z$ has $X|Z \sim Poisson(S, Z)$, then X is called a* ***Cox process*** *conditional on Z.*

Remark For the exact mathematical definition of a **random field** and its properties, see for example Adler (1981). For our purposes it suffices to say that $Z(u)$ is a random variable for all $u \in S$. Hence, we can regard Z as a random intensity function. Under certain regularity conditions, Z can be associated with a corresponding measure Λ, say, i.e. a Cox process can equivalently be defined in terms of a random measure, for details see Møller and Waagepetersen (2003b).

Note that Cox processes are generalisations of Poisson processes in the sense that choosing Z in the above definition to be deterministic yields a Poisson process.

Examples of Cox processes include the **mixed Poisson process**, where the underlying stochastic process Z is expressed in terms of the uniform distribution multiplied by a one-dimensional random variable. This yields a constant but random intensity function λ_0.

Similarly, **random independent thinning** of a Poisson process can be interpreted as a Cox process, where the thinning process is Z. Neyman-Scott processes can also be interpreted as Cox processes, where random allocation to cluster centres is regarded as the underlying random process (Møller and Waagepetersen 2003b).

Other more complex examples of Cox processes comprise the very flexible class of **log Gaussian Cox processes** introduced in Møller et al. (1998), where closed form expressions of second order summary statistics can be found, as well as **shot noise Cox processes** introduced by Brix (1999).

Through conditioning on Z, a number of distributional properties can easily be derived for Cox processes, for details refer to Møller and Waagepetersen (2003a,b).

The hierarchical spatial point process fitted to the locations of reseeding plant species in the Cooljarloo data set in Chapter 6 may be regarded as a (multi-type) Cox process, where the underlying random process is a spatial point process formed by the positions of the resprouters.

3.4.9 Multivariate Cox processes

The definition of a Cox process can be easily generalised to the multivariate case, i.e. a situation where more than one type of points is considered. The generalisation assumes that each of the subprocesses has its own underlying

stochastic process representing unobserved environmental variability; environmental conditions may influence different types of points in a different way. For instance the combined effect of soil type, temperature, moisture and nutrient levels may have a different influence on different plant species and thus influence the pattern formed by each of the species differently.
More formally, we generalise definition 3 as follows:

Definition 4 *Let $Z_i = \{Z_i(\xi) : \xi \in S\}$, $i = 1, \ldots, k$ be non-negative locally finite random fields on S. If the conditional distribution of spatial point processes $X_1, \ldots, X_k$ given $Z = (Z_1, \ldots, Z_k)$, are independent Poisson processes with intensity functions $Z_1, \ldots, Z_k$, then $X = (X_1, \ldots, X_k)$ is called a* ***multivariate Cox process*** *conditional on Z.*

3.4.10 Simulating Cox Processes

As with Markov processes, the simulation of Cox processes other than those which can be regarded as Neyman-Scott processes usually involves the application of Markov Chain Monte Carlo (MCMC) methods, notably in the context of log Gaussian Cox processes and shot noise Cox processes. This usually involves the simulation of the underlying random field Z followed by the simulation of the Poisson process $X|Z$ conditional on this random field. Refer to Møller and Waagepetersen (2003a,b) as well as Berthelsen and Møller (2002) for more details. As mentioned above, we use an approach where the positions of the seeders are modelled conditional on the underlying process Z formed by the resprouting species, in Chapter 6. We will use MCMC methodology to simulate patterns for parameter estimation.

3.5 Parameter estimation

For a given observed spatial point pattern, the underlying spatial point process producing that specific pattern is unknown. It is possible to infer a spatial point process from the observed pattern assuming a specific model for the data. To do so, we apply methods that allow estimation of the parameters of the spatial point process. However, spatial point processes are very complex models and, as mentioned above, the densities are only known for very simple models. As a result, estimation procedures commonly used in more simple statistical approaches, such as maximum likelihood estimation, cannot easily be applied here.

The methods described in Sections 3.5.3 and 3.5.4 are applied in Chapter 6 and are thus described here in greater detail. For more details refer to Møller and Waagepetersen (2003a,b) and the references therein.

3.5.1 Maximum likelihood estimation

Historically, the first approaches developed to estimate the parameters of a spatial point process model from a given pattern applied maximum likelihood estimation (MLE), an approach that seeks to maximise the likelihood of the parameters given the data. While still being approximate, these methods (Ogata and Tanemura 1985) tend to be very computationally intensive (Baddeley and Turner 2000).

An approximate MLE can be found using methods such as **importance sampling** or **path sampling** based on Monte Carlo methods. See Møller and Waagepetersen (2003b) for a detailed treatment of these approaches.

3.5.2 Minimum contrast method

In the context of Neyman-Scott processes, the **minimum contrast method** (Diggle 1983) has become the standard estimation method. Here, parameters are chosen such the differences between the second order summary statistics calculated from the data and the theoretical second order summary statistics are minimised. This can be done since closed form expressions of the theoretical summary statistics are known for some of these models. In this thesis, we do not apply this method as we do not fit a Neyman-Scott process to the data and have only included a short paragraph on this for the sake of completeness. Examples and more details can be found in Diggle (2003), Stoyan and Stoyan (1994) and Møller and Waagepetersen (2003b).

3.5.3 Maximum pseudolikelihood and the Berman Turner device

The estimation method described below can be used for parameter estimation for Markov processes; it will be used in Section 6.1. Furthermore, and more importantly, in Section 6.2 we use a similar but slightly modified approach for the estimation of the parameters in the initial hierarchical model, building on a similar argument as here. In order to make this modification clearer within the context of this thesis, the derivation of the method is explained in greater detail.

The approach described below is mainly based on work by Baddeley and Turner (2000), who extended results by Berman and Turner (1992). It yields a practical method for the estimation of the pseudolikelihood of a

spatial point process that facilitates the estimation procedure of a parametric Markov point process model.

Note that the following sections are of a rather technical nature. For the non-specialist reader it suffices to appreciate that the method described below yields a readily applicable and very stable, yet approximate, approach which may be used for parameter estimation of Markov processes[8].

3.5.3.1 Pseudolikelihood for discrete models

The maximum pseudolikelihood approach was originally introduced outside the context of spatial point processes to overcome the problem of intractable normalising constants by Besag (1975, 1976) in a non-spatial setting. In an analogy to MLE, we seek to maximise the pseudolikelihood of the p parameters $\theta = (\theta_1, \ldots, \theta_p)$ given the data.

Besag (1975) defines the **pseudolikelihood** for random variables $X_1, \ldots, X_n$ as the product of the conditional likelihoods of each X_i given the other variables (Besag 1975, 1976):

$$PL(\theta, x) = \prod_{i=1}^{n} P_\theta\{X_i = x_i | (X_j = x_j, j \neq i)\}$$

and similarly for non-discrete random variables.

For an exponential family model the normalising constant cancels when the conditional likelihoods are considered such that maximisation of the pseudolikelihood by minimising the logarithm of the reciprocal can be easily computed (Baddeley and Turner 2000). Furthermore, if the random variables

[8]It will become apparent from what follows that the model has to be of exponential form for the method to work. Since this is not the case for the Cox process models described in Section 3.4.8 different estimation methods have to be used in that context.

$X_1, \ldots, X_n$ are only weakly dependent, the pseudolikelihood yields a good approximation to the likelihood and MPLE estimators are unbiased, consistent and asymptotically normal under certain conditions (Jensen and Møller 1991).

3.5.3.2 The pseudolikelihood approach for spatial point patterns

As mentioned above, for point processes other than the inhomogeneous Poisson process, it is difficult to maximise the likelihood. Even simple models include a normalising constant, which is an intractable function of the parameter vector θ. Thus, the pseudolikelihood approach was extended to spatial point processes by Besag (1976). It can be regarded as an infinite product of infinitesimal conditional probabilities.

In analogy to the approach taken with general random variables, where conditional probabilities were considered, we introduce the notion of the conditional intensity in order to construct the pseudolikelihood for spatial point processes; for more details on conditional intensities refer to Kallenberg (1984). The **conditional intensity** of any Markov process on S with density f is defined as

$$\lambda(u; x) = \frac{f(x \cup \{u\})}{f(x)}, \quad u \notin X, \tag{3.11}$$

$$\lambda(\xi; x) = \frac{f(x)}{f(x \backslash \{\xi\})}, \quad \xi \in X. \tag{3.12}$$

Note that the intractable normalising constant in for instance (3.10) may be eliminated in the conditional density.

Examples

For the homogeneous Poisson process

$$\lambda(u; x) = \lambda$$

and the inhomogeneous Poisson process yields

$$\lambda(u; x) = \lambda(u).$$

For a general Markov process $\lambda_\theta(u; x)$ depends on x.

For example, the pairwise interaction process has conditional intensity

$$\lambda(u; x) = \beta(u) \prod_{\xi,\eta \in X}^{\neq} h(u, \xi)$$

Now, Besag (1976) defines the **pseudolikelihood of a point process** with conditional intensity $\lambda(u; x)$ over S as

$$PL_S(x) = \left(\prod_{\xi \in X} \lambda(\xi; x) \right) \exp\left(- \int_S \lambda(u; x) du \right). \tag{3.13}$$

Examples

If the process is Poisson the pseudolikelihood coincides with the likelihood up to a factor $\exp(|S|)$.

For a pairwise interaction process, the pseudolikelihood is

$$PL_S(\theta; x) = \left(\prod_{i=1}^{n(x)} \beta(\xi) \right) \left(\prod_{\xi,\eta \in X}^{\neq} h(\xi, \eta) \right) \exp\left(- \int_S b(u) \prod_{i=1}^{n(x)} h_\theta(u, \xi) du \right).$$

3.5.3.3 Loglinear case

In the following, we focus on Markov point process models where the conditional intensity is loglinear, i.e. where it can be written as

$$\lambda(u; x) = \exp(\theta' C(u; x)).$$

Here, $C(u; x)$ is a vector of spatial covariates defined at each point u in S. Assume $||C(u; x)|| \exp(\theta' C(u; x))$ is uniformly bounded in $u \in S$ and $\theta \in \Theta$ for each fixed x. Then the maximum pseudolikelihood normal equations

$$\frac{\partial}{\partial \theta} \log PL_A(\theta; x) = 0$$

become

$$\sum_{\xi \in x} C(\xi; x) = \int_S C(u; x) \exp(\theta' C(u; x)) du. \tag{3.14}$$

The solution of the above equation usually requires iterative algorithms. Jensen and Møller (1991) show that (3.14) is an unbiased estimation equation, i.e. that the expectation of both sides of the equation are equal. The MPLE is again consistent and asymptotically normal under certain regularity conditions regarding finiteness and boundedness of the process that are typically satisfied by data realisations.

3.5.3.4 Practical Estimation

Berman and Turner (1992) introduced a parsimonious method which they use to estimate the maximum likelihood for an inhomogeneous Poisson point process. Here, we describe a generalisation due to Baddeley and Turner (2000). They approximate the integral in (3.13) by a finite sum and show

that MPLE is then equivalent to maximizing the likelihood of a generalised linear model, in particular a Poisson model.

Let x a realisation of a Markov process X with conditional intensity $\lambda(u;x)$ as defined above. Consider the pseudolikelihood and approximate the integral in (3.13) by a finite sum using any quadrature rule

$$\int_S \lambda(u;x)du \approx \sum_{j=1}^{m} \lambda(u;x)w_j,$$

where $u_j, j = 1,\ldots,m$, are locations in S and $w_j > 0$ are quadrature weights summing to $|S|$. The u_j are either data points or locations in S that are not data points (termed "dummy points"). The log pseudolikelihood can be approximated as

$$\log PL(\theta;x) \approx \sum_{\xi\in x} \log\lambda(\xi;x) - \sum_{j=1}^{m} \lambda(u_j;x)w_j.$$

If the list of points $\{u_j, j = 1,\ldots,m\}$ contains all the data points $\{\xi \in x\}$, we can rewrite the above equation as

$$\log PL(\theta;x) \approx \sum_{j=1}^{m} (y_i \log\lambda_j - \lambda_j)w_j, \tag{3.15}$$

where $\lambda_j = \lambda_\theta(u_j)$ and $y_j = z_j/w_j$, and

$$z_j = \begin{cases} 1 & \text{if } u_j \text{ is a data point, } u_j \in x \\ 0 & \text{if } u_j \text{ is a dummy point, } u_j \notin x. \end{cases} \tag{3.16}$$

Note that the right hand side of (3.15) is equivalent to the log-likelihood of independent Poisson variables Y_k with means λ_k taken with weights w_k. Thus (3.15) may be maximised using standard software for fitting generalised linear models such as GLIM, S-PLUS or R. This is particularly useful since

it yields a greater stability than maximum likelihood approaches based on purpose-written code. Furthermore, applying this approach will facilitate the estimation procedure and thus make it available for use to non-specialists.

Procedure

1. Generate a set of dummy points, and combine it with the data points x_i to form the set of quadrature points u_i.

2. Compute the quadrature weights w_j.

3. Form the indicators z_j as in (3.16) and calculate $y_j = z_j/w_j$.

4. Compute the values $v_j = S(u_j; x)$ of the sufficient statistic at each quadrature point.

5. Invoke the model-fitting software, specifying that the model is a Poisson regression with log link

$$\log \lambda_j = \theta' v_j$$

to be fitted to the responses y_j and covariate values v_j with weights w_j.

Note that the estimates of the standard error returned by the software are not applicable directly since they assume independent identically distributed Poisson observations and may only be used to construct approximate confidence intervals.

3.5.3.5 Example: The Strauss process

As an illustration we consider the simple **Strauss process model**. Here the conditional intensity is

$$\lambda(u) = \beta\gamma^{t(u,x)},$$

where

$$t(u,x) = \sum_{\xi \in x} \mathbf{1}\{\xi \in x : 0 < ||\xi - u|| \leq r\} \tag{3.17}$$

is the number of points $\xi \in x$ which lie r units or less apart from u. The pseudolikelihood is

$$PL(\beta, \gamma; x) = \beta^{n(x)}\gamma^{s(x)} \exp\left(-\beta \int_W \gamma^{t(u,x)} du\right),$$

i.e.,

$$\begin{aligned} \lambda(u;x) &= \beta\gamma^{t(u,x)} \\ \Leftrightarrow \quad \log(\lambda(u;x)) &= \log\beta + t(u,x)\log\gamma. \end{aligned}$$

This is the required log-linear form with $\theta = (\log\beta, \log\gamma)'$ and $S(u;x) = (1, t(u,x))'$. Hence, to compute the approximate MPLE we fit the following loglinear model

$$\log\lambda_j = \theta_1 + \theta_2 v_j, \tag{3.18}$$

where $v_j = t(u_j;x)$ with $t(u_j;x)$ as defined in (3.17). In S-PLUS or R this is equivalent to the following expression, using the function glm for generalised linear models:

```
glm(y~v, family= poisson, link=log, weights = w)
```

where y, v, w, are S-PLUS vectors of equal length containing the responses y_j, the "explanatory values" v_j and the weights w_j respectively, for each quadrature point u_j.

3.5.4 MCMC for parameter estimation in a Bayesian context

The Bayesian statistical approach follows a slightly different philosophy from the classical or frequentist approach. It assumes that prior information on model parameters in the form of a distribution is available. This prior information together with the likelihood based on the data, eventually yields a posterior distribution of the parameters, given the data. Overall, parameter estimation in Bayesian models is more flexible, which allows for more complex models to be fitted, a fact that is particularly useful in the context of spatial point process modelling.

Previously, it has been impossible to evaluate the likelihood of many complex Bayesian models necessary for parameter estimation. The advent of MCMC has made it possible to circumvent this problem. In this thesis, we will use a Bayesian modelling and estimation approach when fitting a multivariate Cox process to the Cooljarloo data in Chapter 6. Refer to Møller and Waagepetersen (2003b) for a detailed review of MCMC methodology in the context of spatial point processes and in particular in a Bayesian setting.

3.6 Model validation using simulations

Once a spatial point process model has been fitted to a spatial point pattern its fit has to be assessed in order to validate the model and its appropriateness

for the given pattern. This is commonly done by comparing second order summary statistics of a number of simulations (typically 39) of the model to the estimated second order summary statistics for the data[9].

Simulation envelopes are constructed by determining the minimum and the maximum value of the summary statistics for the simulated patterns at a given distance r. These form the upper and lower envelopes and are then plotted along with the estimated summary statistic for the data. The top line of the envelope is the maximum of the 39 simulations at each distance and the bottom line is the minimum. If the L-function for the species of interest is above the top line of the envelope we have a clustered pattern and if it is below the bottom line we have a regular pattern. See Figure 3.11 for an example of the estimated inhomogeneous L-function with simulation envelopes derived from 39 simulations of an inhomogeneous Poisson process for the species *Leucopogon striatus*.

In addition, a further line has been added to the plot which should be close to zero for most distances. This is the mean of the L-functions of the 39 simulated patterns at each distance. It has been included to verify whether the simulations were indeed homogeneous Poisson patterns themselves. It moves away from the zero line at larger distances resulting from an estimation bias for these distances.

The fact that the estimated L-function for the data is above the upper envelope for smaller distances indicates a clustered pattern that cannot be explained by the simple inhomogeneous Poisson model applied here.

[9] 39 simulations are traditionally chosen as this is the smallest number of simulations necessary detect significant differences at a 5% level.

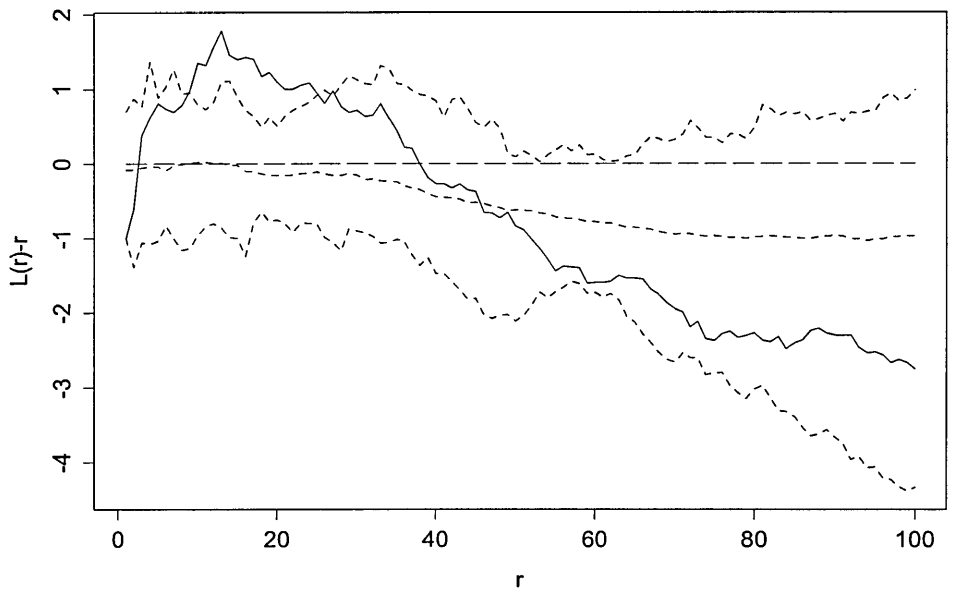

Figure 3.11: Estimated inhomogeneous L-function with simulation envelopes for species *Leucopogon striatus*. The dashed lines are the upper simulation envelope, mean of the L-functions at each distance and lower simulation envelope respectively. The full line is the estimated L-function for the data.

In Chapter 4, plots like the one in Figure 3.11 will be used extensively for an initial analysis of the data and a classification of the patterns for each of the species in groups of similar spatial behaviour.

3.7 Outlook

3.7.1 Spatial point process modelling in ecology

Spatial point process models have rarely been applied in ecology. Most of the time the applications are typically merely descriptive, only making use of first and second order summary statistics (Perry et al. 2002). The few

applications in the literature are restricted to small numbers of species – at most three – but have never been applied to model a plant community, possibly due to the complexity of the necessary models but also due to most ecologists' unawareness of the methods available.

Within the statistical literature, most applications of spatial point processes so far have been theory driven rather than data driven. This has led to the development of potential models which might be applied to data sets, given they are available. The standard methodology for fitting and evaluating a spatial point process model is highly computationally intensive. As a result, if at all, these methods were applied to data sets with a very small number of species – both for ease of computation and due to a lack of more complex data sets.

3.7.2 Approaches taken in this thesis

Data sets in ecology, however, especially those of communities with high biodiversity, tend by definition to be much more complex. With 6378 plants from 67 species on a 22m $\times$ 22m plot, the study data set consists of an unusually large number of plants from a similarly unusually large number of species. In order to make spatial point processes applicable in this context, appropriate novel methods have thus to be developed to deal with this situation to yield a methodology capable of handling large data sets and to **reduce the complexity** of the data set. In Chapter 4, we introduce a method that will allow us to filter out – in a parsimonious way – the patterns of those species which apparently deviate from complete spatial randomness. In Chapter 5, we present a novel method, which applies functional principal

component methods to second order summary statistics in order to derive a principal component method for spatial point processes. This will enable us to reveal the main groups of species with similar spatial characteristics and their similarity and thus reduce the complexity of the data set.

Furthermore, so far, spatial point process models, such as the multi-type processes described in Section 3.4.4.3, have typically assumed symmetric inter-species interaction. This assumes that, when two species interact, there is mutual attraction or repulsion. Typically, this is not realistic in the context of plant community ecology (Watkinson 1997). In this application, for instance, we initially fit a Continuum Ising model to the data, only to show that it is inappropriate in this context as it ignores asymmetric interaction structures.

Subsequently, we acknowledge that the different regeneration methods as described in Section 2.4.4.1 give rise to an asymmetric interaction structure, which has to be reflected in any suitable model. Thus, we develop a **hierarchical** spatial point process model, more specifically a Cox process, where the underlying environmental process is formed by the resprouting species whose location has been fixed for thousands of years. This is more appropriate in this context in particular, but more generally will be applicable to other plant communities and even to spatial point patterns formed by other objects, where an asymmetric interaction structure between the objects has to be assumed.

Chapter 4

Exploratory analysis of data set

This chapter commences the formal analysis of the Cooljarloo data set, introduced in Section 2.4, by exploring the first and second order characteristics of the subpatterns[1]. Through this, we are able to identify those subpatterns that deviate from complete spatial randomness and this consequently informs the modelling process.

The usual initial step in the analysis of a spatial point pattern is an exploratory data analysis consisting of a careful visual inspection of the pattern. This is followed by the investigation of the first and second order summary statistics as introduced in Section 3.3. This will inform the next steps, i.e. the choice of an appropriate model and parameter estimation.

In the case of the given pattern derived from the Cooljarloo data set, the exploratory analysis turns out to be a rather time consuming process given the large number of subpatterns and the resulting large number of plots, summary statistics and intra- and inter-species interactions. The pre-

[1]In this thesis the term "subpattern" refers to the spatial point pattern formed by one particular species. The Cooljarloo data set consists of the locations of 67 species so consists of 67 subpatterns, some of which contain only one point.

sentation in this chapter will thus be restricted to a number of interesting and relevant examples based on species which we will focus on in subsequent chapters. The plots of all patterns and all summary statistics with simulation envelopes as well as complete tables summarising the properties of all those species where an analysis of the pattern was deemed meaningful may be found in Appendix B.

We start by applying first order summary statistics to the subpatterns, by displaying the estimated intensity surfaces, see Section 4.1. We then introduce a new approach to formal statistical testing for inhomogeneity in the pattern based on a permutation approach that avoids some of the problems faced with existing techniques. This mainly involves the fact that a single subdivision of the plot has to be chosen *a priori*. We also present a further new approach based on the one-sample Kolmogorov-Smirnov test (D'Agostino and Stephens 1986) that assesses the inhomogeneity of a pattern in x-direction or in y-direction or both, and so considers the direction of an inhomogeneity or trend.

Section 4.2 presents the application of second order summary statistics to the subpatterns and points out second order characteristics typical of the patterns in this data set. In Section 4.3 we provide a summary of the results obtained so far. In Section 4.4 we present a novel, parsimonious approach to provide a formal test of complete spatial randomness which turns out to be particularly useful in the context of a highly multi-variate pattern such as the current data set. We present the results from a simulation study which was used to assess the statistical power of the new approach and close with a discussion of methods and results.

4.1 First order summary statistics

4.1.1 Intensity estimation

The first step in the analysis of the data set consists of assessing first order properties, i.e. estimating the intensity surface for each of the subpatterns[2] and investigating whether the patterns are homogeneous or inhomogeneous. Information on the homogeneity and inhomogeneity of the pattern is useful for various reasons. When second-order summary statistics are applied (see Section 4.2), more specifically the K- and L-functions, the inhomogeneous versions of these have to be chosen instead of the simple homogeneous version. Furthermore, a model that describes these patterns must explicitly incorporate the inhomogeneity.

Finally, the first order characteristics may yield valuable biological or ecological information revealing properties of unobserved environmental variability that is reflected in the pattern. For instance, the inhomogeneity of a pattern formed by plants might be due to an inhomogeneity of nutrients in the soil that researchers have previously not been aware of or are unable to measure. Similarly, the analysis may give an initial indication of larger scale interspecies interactions. Areas of low intensity for one species and high intensity of another species might also indicate larger scale interaction between the species.

[2]Note that in this chapter, we only include those patterns with number of points ≥ 20, since we consider patterns with a smaller number of points too sparse to yield interpretable information on their spatial formation. In the given data set this excluded 32 species out of 67 from the formal analysis. However, we shall see in Chapter 6 that the sparse species may still be included in a model as they may influence the pattern of other less sparse species.

We present here only some examples of intensity surfaces with behaviour typical of the subpatterns in the data set[3]. Figure 4.1 shows plots of the estimated intensity surfaces for the species *Conostephium pendulum* (a) and *Scholtzia involucrata* (b). It is clear that the two processes are very likely to be inhomogeneous. *Conostephium pendulum* shows a low intensity towards the western end of the plot and a higher density of plants at the eastern end. *Scholtzia involucrata* shows a clear north-south trend with a higher intensity at the southern end of the plot.

These results indicate that investigating the potential inhomogeneity may indeed be an issue with this data set such that in further steps of analysis this has to be taken into account and homogeneity cannot be assumed. Appendix B shows the estimated intensity surfaces for all species. A careful inspection of these estimated surfaces reveals that a large proportion of the subpatterns appears inhomogeneous. In particular, both a trend in north-south direction and a trend east-west direction occur repeatedly in the data set. However, no direct explanation for this could be found (Armstrong 2005).

In most applications in the literature, the first order characteristics of a spatial point pattern have merely been assessed by the visual inspection of the estimated intensity surfaces (Stoyan et al. 1995) and formal tests of inhomogeneity have neither been frequently used nor discussed in the literature. However, the informal inspection of the plots can only give an indication of inhomogeneity in the patterns and does not yield an objective decision

[3]Note that for the estimated surfaces presented here no edge correction was done leading to apparent edge effects. Alternatively, edge correction could have been applied, as described in Section 3.3.3. This may lead to spurious clustering for inhomogeneous patterns or loss of data, depending on the edge correction method used.

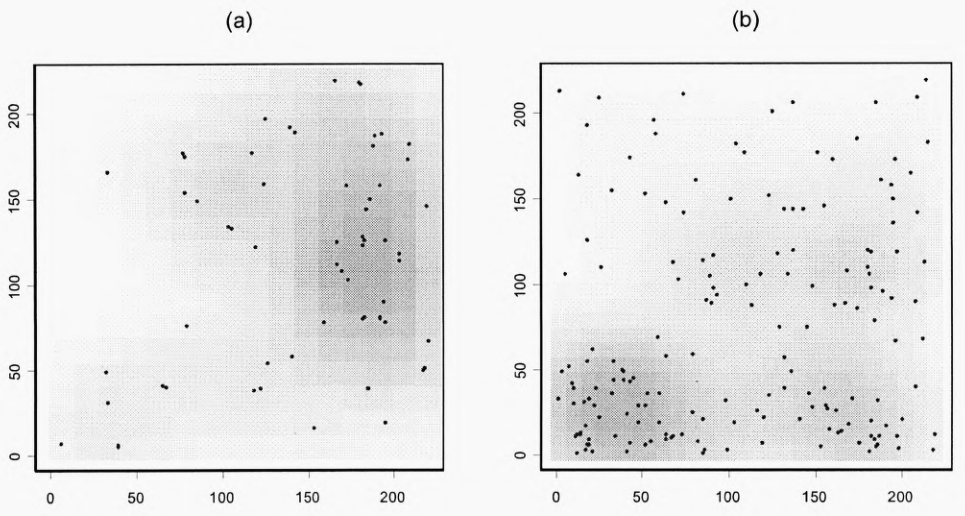

Figure 4.1: Estimated intensity surfaces for *Conostephium pendulum* and *Scholtzia involucrata*; measurement unit is 10cm.

criterion. For this reason, we assess inhomogeneity using formal hypothesis testing in Section 4.1.2 and suggest some new approaches to this.

4.1.2 Testing for inhomogeneity

4.1.2.1 Permutation test

The formal test for inhomogeneity described in Section 3.3.5 depends heavily on the choice of the subareas of the plot. Strictly speaking, this choice has to be made *a priori* since the subdivision has to be independent of the actual pattern. The choice may be made either at random or based on background information, but should not be based on an inspection of the pattern itself (Stoyan and Stoyan 1994). Furthermore, depending on the choice of the subdivision, not all types of inhomogeneity might be discovered.

As an example, consider the simulated pattern in Figure 4.2, which has been generated from an inhomogeneous Poisson process. If this pattern is subdivided as indicated by the line in the plot, the apparent inhomogeneity will not be detected by the test; using the test statistic in (3.8) and the suggested subdivision, yields $F(128, 52) = 0.948, p = 0.604$ providing no evidence against the pattern's homogeneity. Similar results can be expected from most other choices of a horizontal line for this specific pattern.

As a consequence, we suggest a **permutation test** (Good 2000) where more than one subdivision is considered. Here, the plot area is subdivided into two areas n_{perm} times, by randomly finding the position of either a vertical or horizontal line with equal probability each time[4]. The test statistic intro-

[4]A further generalisation of the approach could consider more general subdivisions of the area, by also randomly choosing an angle for the subdivision, where the angle is chosen from a uniform distribution between 0 and 90 degrees. This thesis will not consider this

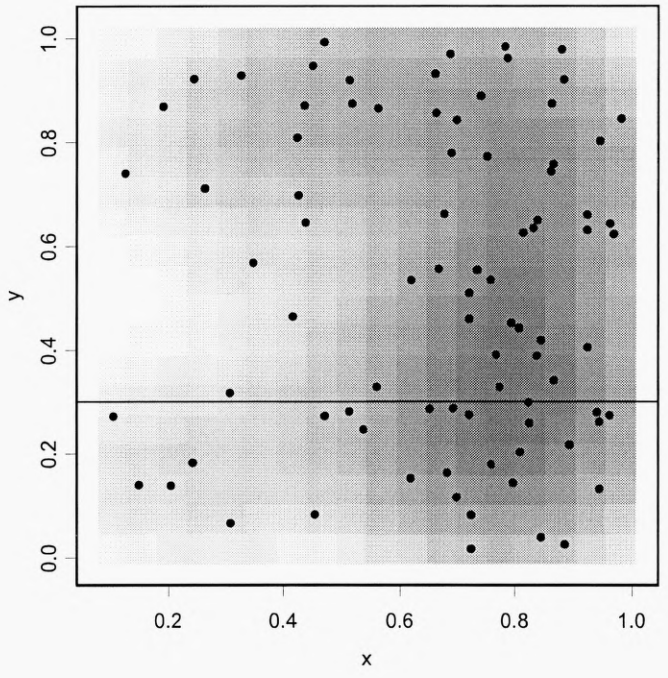

Figure 4.2: Realisation of a pattern from an inhomogeneous Poisson process with a trend in x-direction and horizontal subdivision.

duced in equation (3.8) in Section 3.3.5 is then calculated for the resulting subdivisions. Subsequently, the mean test statistic is calculated from the n_{perm} test statistics to yield a single test statistic value.

It remains to determine whether the resulting test statistic is likely to have occurred under the null hypothesis of homogeneity. However, the distribution of this test statistic is not known, so we suggest an approach where the

generalisation. In this chapter we aim at pointing out the fact that there is ample scope for improving the existing test for inhomogeneity in particular if no prior knowledge on potential inhomogeneity is available. We present two simple generalisations of the F-test.

distribution is estimated from simulated patterns. Thus, n_{sim} homogeneous Poisson patterns with expected number of points as the point pattern under investigation are simulated and each of these is treated in the same way as the observed pattern. This yields a distribution of the values of the test statistic for homogeneous Poisson patterns, i.e. the distribution of the statistic under the null hypothesis. The results for the simulated pattern are compared to those for the empirical pattern to assess the significance level.

4.1.2.2 Kolmogorov Smirnov approach

The permutation test remains a rather coarse approach in the sense that the plot area is subdivided into two subsections only. Even though this is repeated n_{perm} times, it is likely that this leads to a low power of the test as ill-positioned divisions will repeatedly not detect an apparent inhomogeneity. Essentially, the permutation test, like the original test, tests for trends in the x- and the y- direction at the same time, resulting in a potential decrease in power. For instance, if the observed pattern shows a trend in x-direction, approximately 50% of the simulated patterns would still be tested for a trend in y-direction, which is likely to further decrease the power of the test. Furthermore, a significant result of the test would still not indicate the nature of the trend, i.e. whether there is a trend in x- or y-direction.

We therefore suggest an alternative approach which separately considers the one-dimensional distributions of the x- and y-coordinates, respectively, and then tests these for a uniform distribution, using a Kolmogorov-Smirnov goodness of fit test (see, e.g. D'Agostino and Stephens (1986)). If a pattern

were homogeneous, the one-dimensional distributions should follow a uniform distribution.

To illustrate this, Figure 4.3 shows the pattern formed by species *Astroloma xerophyllum*, the estimated intensity surface as well as histograms for the x-and y-coordinates of the pattern. If the pattern were homogeneous, the histograms of the empirical distributions would be close to uniform. It is apparent that this may not be the case here for the y-coordinates, given that there are no values above 165, i.e. no plants in the northern 5.5 metres of the plot[5]. The Kolmogorov Smirnov test confirms this, indicating no evidence of inhomogeneity in x-direction ($p > 0.05$) and inhomogeneity in y-direction ($p < 0.001$).

However, this is a multiple-testing problem so we adjust the significance level in the following way. The test statistic of the Kolmogorov Smirnov test in x-direction solely depends on the x-coordinates of the pattern and the test statistic of the Kolmogorov Smirnov test in y-direction on the y-coordinates. Under the null hypothesis ("The pattern is a realisation of a homogeneous Poisson process") the x- and the y-coordinates are independent and so are the two test statistics. Let c_α be the $1-\alpha$ quantile of the distribution of the Kolmogorov Smirnov test in x-direction (KSx), such that

$$P(\text{ KSx } > c_\alpha) = \alpha$$

[5]Measurements have been made in units of 10cm.

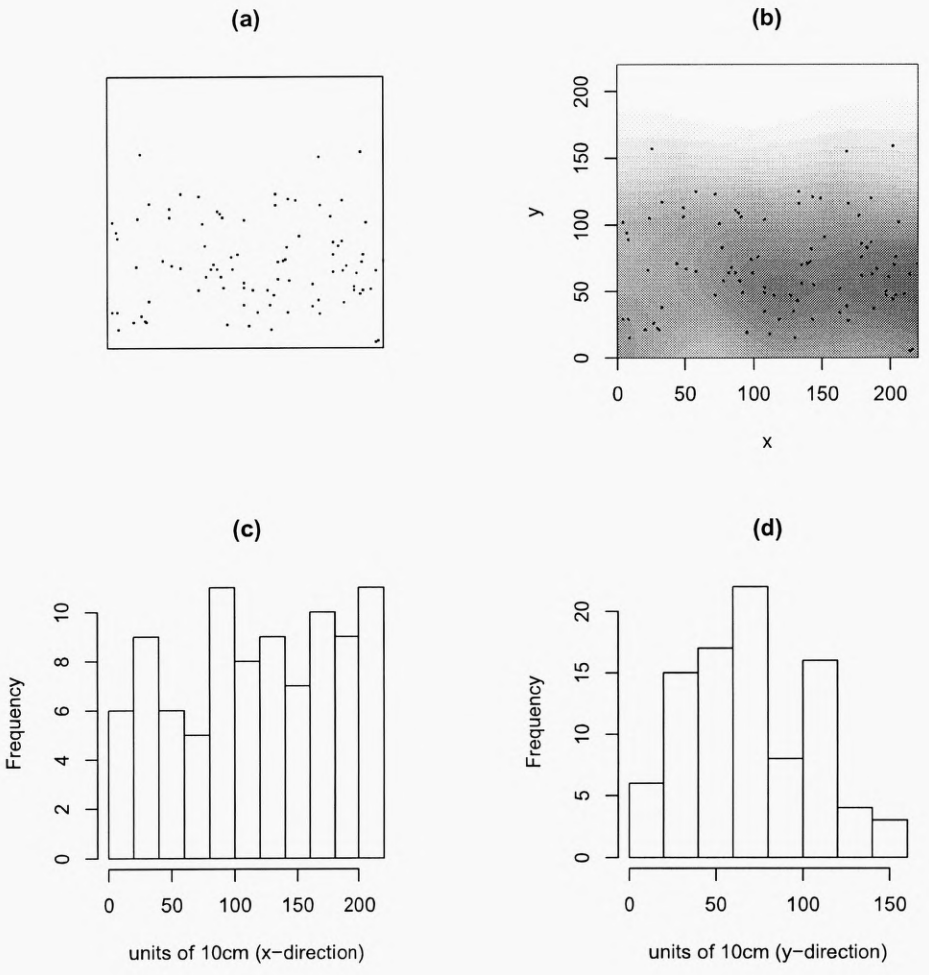

Figure 4.3: Pattern, estimated intensity surface and histograms for x- and y-coordinates of species *Astroloma xerophyllum*; measurements in units of 10cm.

and similarly for KSy. For the multiple testing problem we have

$$\begin{aligned} 1 - P(\text{ KSx } \leq c_\alpha,\ \text{KSy } \leq c_\alpha) &= 0.05 \\ \Leftrightarrow \quad 1 - (1-\alpha)^2 &= 0.05, \end{aligned}$$

which yields $\alpha = 1 - \sqrt{0.95} = 0.0253$. Thus the Kolmogorov Smirnov test is considered significant if either one or both of the p-values for KSx or KSy are smaller than 0.0253 [6].

We apply all three approaches, the original F-test, the permutation test and the Kolmogorov Smirnov test to the data set and then compare the results.

4.1.2.3 Results

Table 4.1 presents the results for some selected species. These include species for which the results serve as illustrations of typical characteristics of the patterns in the dataset and of the typical differences in performance for the three methods. We also include species, which will be considered in the model in Chapter 6.

The table lists the p-values for the F-test described in Section 3.3.5, the permutation test and the Kolmogorov Smirnov approach suggested here. For the F-test the subdivision was chosen *a priori* at random and separately for each species as no background information was available as to a potential underlying inhomogeneity in environmental conditions that may have led to an inhomogeneity with a specific trend in the data. This was done twice – once in x-direction and once in y-direction. The α-level was adjusted accordingly, as described in Section 4.1.2.2. Refer to Tables B.1 and B.2 in Appendix B for a list of the results for all species with more than 20 plants.

[6]Note that this yields a single decision criterion at the 5% level as to whether there is evidence that a subpattern may be considered inhomogeneous. Strictly speaking, there is also an issue of multiple testing across species which we do not deal with any further at this exploratory stage of the analysis

species	F-test	Permutation test	Kolmogorov-Smirnov x-direction y-direction
Andersonia heterophylla	$p = 0.967$	**p = 0.000**	**p = 0.004** **p = 0.003**
Astroloma xerophyllum	**p = 0.001**	**p = 0.000**	$p = 0.36$ **p = 0.000**
Banksia attenuata	$p = 0.734$	$p = 0.71$	$p = 0.72$ $p = 0.27$
Bossiaea eriocarpa	$p = 0.497$	$p = 0.11$	$p = 0.90$ $p = 0.0375$
Conospermum crassinervium	$p = 0.274$	$p = 0.23$	$p = 0.08$ $p = 0.0487$
Conostephium pendulum	$p = 0.730$	$p = 0.59$	**p = 0.000** $p = 0.889$
Jacksonia floribunda	$p = 0.982$	**p = 0.000**	$p = 0.0676$ $p = 0.0322$
Leucopogon conostephioides	$p = 0.265$	$p = 0.41$	**p = 0.0234** **p = 0.0103**
Leucopogon striatus	$p = 0.914$	**p = 0.000**	$p = 0.131$ **p = 0.000**
Scholtzia involucrata	$p = 0.989$	$p = 0.67$	$p = 0.45$ **p = 0.000**

Table 4.1: p-values for three tests of inhomogeneity of patterns for selected species; significant p-values are in bold

A comparison of the different p-values in Table 4.1 and in Tables B.1 and B.2 in Appendix B clearly indicates the strengths and weaknesses of the different approaches. There were a number of cases where both the F-test and the permutation test did not detect any inhomogeneity but the Kolmogorov Smirnov test did, but there are only two cases where the Kolmogorov Smirnov test did not detect an inhomogeneity which the F-test or the permutation test detected. This might indicate that the Kolmogorov Smirnov test is bet-

ter at detecting an existing inhomogeneity but might also indicate that the Kolmogorov Smirnov test is too sensitive and classifies patterns as inhomogeneous even if they are not. A simulation study in 4.1.2.4 investigates this issue in more detail and we shall now refer to some examples from the data set for illustration.

For instance, for species *Conostephium pendulum* the F-test yields a non-significant p-value of 0.73 and the permutation test yields a non-significant p-value of 0.59, whereas the Kolmogorov Smirnov test indicates a highly significant inhomogeneity in the x-direction. An inspection of the plots in Figure 4.4 indicates that the Kolmogorov Smirnov test has been more sensitive in detecting the inhomogeneity in the x-direction, which is clear from the estimated intensity surface (Figure 4.4 (b)) and the x-direction histogram (Figure 4.4 (c)).

Similarly, for species *Scholtzia involucrata*, the F-test yields a p-value of 0.989, the permutation test 0.67, whereas the Kolmogorov Smirnov test indicates inhomogeneity in y-direction (p= 0.000), which is consistent with the plots in Figure 4.5.

The results for *Andersonia heterophylla* provide an extreme example of the difference in sensitivity of the three approaches. The F-test yields a very large p-value (0.967), whereas both the permutation test and the Kolmogorov Smirnov test indicate strong inhomogeneity, the latter even indicates that there is inhomogeneity in both x-direction and y-direction. Figure 4.6 shows that the point density for the species is rather low in the top right hand corner of the plot. This shows as a slight decrease in the histogram in x-direction and a stronger decrease in y-direction. It remains to be investigated

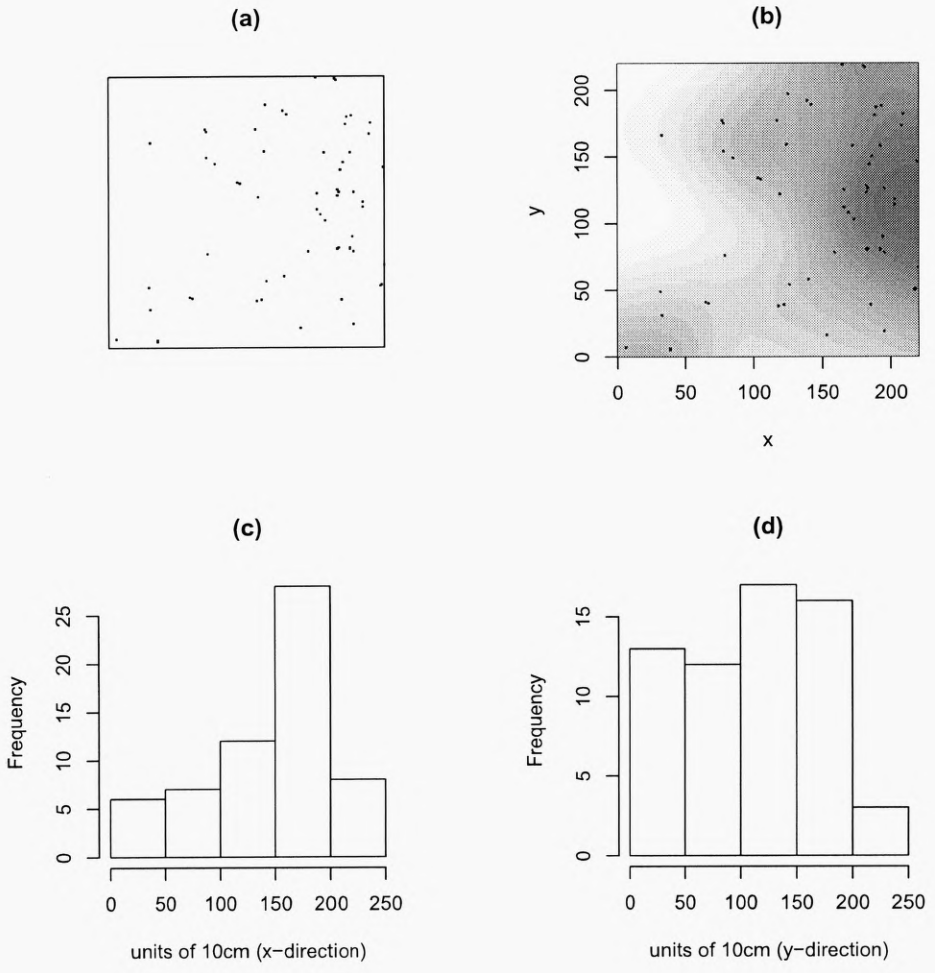

Figure 4.4: Pattern, estimated intensity surface and histograms for x- and y-coordinates of species *Conostephium pendulum*.

whether this is indicating that the Kolmogorov Smirnov test is too sensitive, in particular when considering the only slight decrease in x-direction. The simulation study in Section 4.1.2.4 will provide further insight into this issue. *Astroloma xerophyllum* is one of the few cases where all three tests agree that the pattern is inhomogeneous, the cases of *Banksia attenuata*, *Bossiaea erio-*

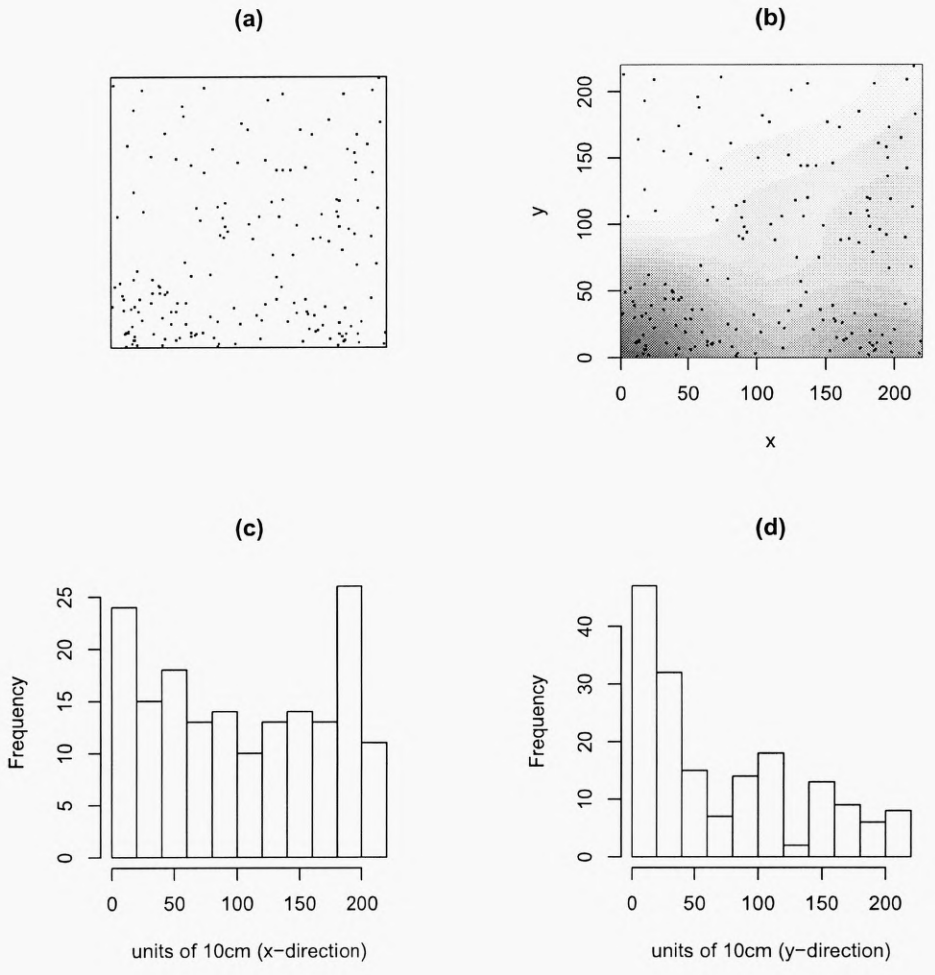

Figure 4.5: Pattern, estimated intensity surface and histograms for x- and y-coordinates of species *Scholtzia involucrata*

carpa and *Conospermum crassinervum* provide other examples of agreement of the three tests, but this time for homogeneity. The results for *Conostephium pendulum* demonstrate that the Kolmogorov Smirnov test may be more sensitive than the other two- most notably the F-test. It indicates inhomogeneity whereas neither the F-test nor the permutation test detect

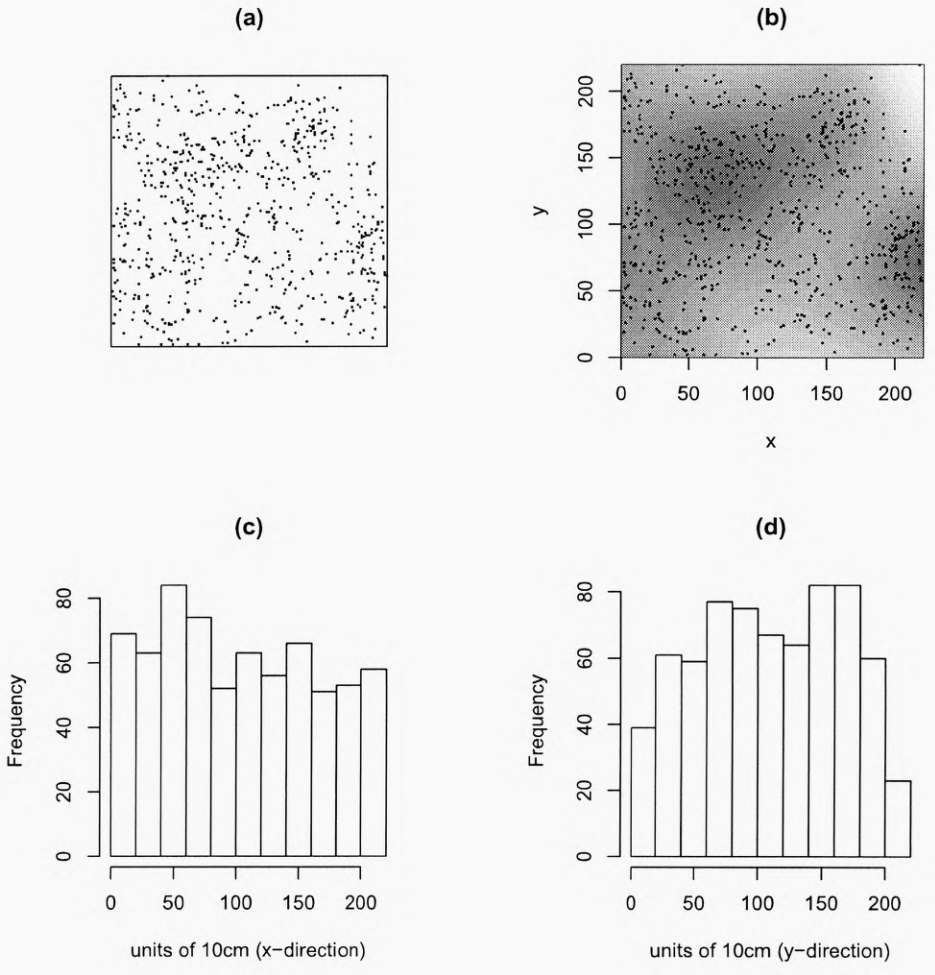

Figure 4.6: Pattern, estimated intensity surface and histograms for x- and y-coordinates of species *Andersonia heterophylla*

inhomogeneity for the pattern. The species *Jacksonia floribunda* is an interesting example where both the F-test and the Kolmogorov Smirnov test are not significant but the permutation test is. However, there is some indication that the Kolmogorov Smirnov test is more sensitive than the F-test with the latter resulting in a p-value of 0.982 and the Kolmogorov Smirnov test resul-

ting in a p-value of 0.0322 which is much closer to the significance level of 0.0253. The species *Leucopogon conostephioides* provides another example of the sensitivity of the Kolmogorov Smirnov test which is the only one of the three tests to reveal the pattern's inhomogeneity. For the species *Leucopogon striatus*, both the permutation test and the Kolmogorov Smirnov test find inhomogeneity, but the F-test does not.

4.1.2.4 Simulation study

On the whole, the above results demonstrate that the Kolmogorov Smirnov approach may be more sensitive than the other two approaches and thus have more power when detecting deviations from the null hypothesis of homogeneity. To this end, a simulation study was run, to understand better the power of the Kolmogorov Smirnov approach in comparison to the classical F-test in different situations where the properties of the patterns are known[7]. Here power is defined as the probability of the test rejecting the null hypothesis for a given set of alternative hypotheses. Inhomogeneous Poisson patterns were generated with a trend function $f(x, y) = a_{sim} \cdot x$ where the value of a_{sim} was gradually increased to mimic an increasing trend in x-direction. Note that this implies that with increasing a_{sim} the simulated point patterns consist of a larger number of points.

For each value of a_{sim} 100 patterns were simulated and both the F-test and the Kolmogorov Smirnov test were calculated[8].

[7]The permutation test was not included in this study due to infeasibly long running times.

[8]Both tests were run for the $x-$ and the y-direction, using an α-adjustment as described above, as if the direction of the trend was not known. This was done to ensure that the

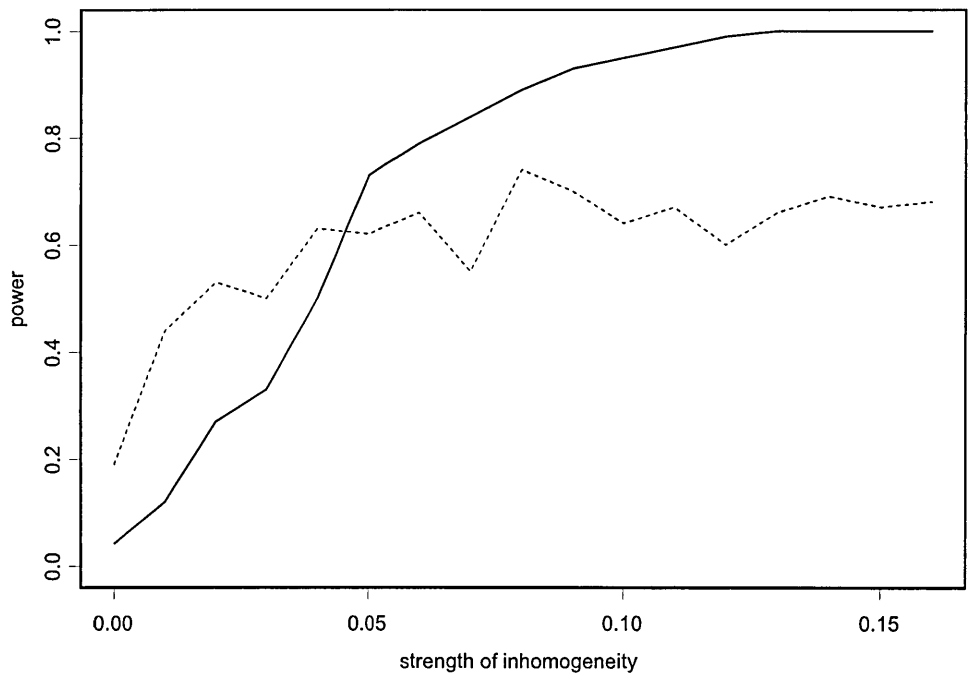

Figure 4.7: Results from power study using inhomogeneous Poisson patterns with increasing linear trend; the full line shows the power of the Kolmogorov Smirnov approach and the dashed line the power of the F-test

Figure 4.7 shows the results of the simulation study. The estimated power of the F-test is apparently higher for low degrees of inhomogeneity than that of the Kolmogorov Smirnov test for weak inhomogeneity but does not substantially improve with increasing inhomogeneity. The power of the Kolmogorov Smirnov test, however, while being not very powerful initially, quickly rises to reach a power of one.

---

tests are compared based on their different approaches. Had the F-test been done for a random direction only, it would have been wrong in 50% of cases, leading to a decrease in power unrelated to the performance of the approach. Applying the test in x-direction only would have yielded the same results.

An explanation for the higher estimated power of the F-test for lower degrees of inhomogeneity compared to the Kolmogorov Smirnov approach may be found in the specific choice of alternatives chosen for this study. However, it also finds (spurious) inhomogeneity for 19% of the homogeneous patterns (strength of inhomogeneity equals 0) whereas the Kolmogorov Smirnov approach only identifies 4% of these patterns as inhomogeneous, which is close to the 5% alpha level. Further, if different types of inhomogeneity had been assessed, the F-test would have had more problems to detect this. Figure 4.8 shows an example of a pattern with Gaussian trend ($\mu = 0.5, \sigma = 2$) in x-direction. An inhomogeneity of this type should be not more difficult to detect for the Kolmogorov Smirnov test than the linear inhomogeneity described above. A formal power study would have varied the value of σ to mimic variation in strength of inhomogeneity. To maintain an equal number of points across patterns these would have to be normalised by dividing by the integral of the trend function over the area. The power in this specific case was 0.69 for the F-test and 1 for the Kolmogorov Smirnov test.

Other goodness of fit approaches, such as a chi-squared approach, may be applied and their performance may be compared to the approaches described here. This is beyond the remit of this thesis.

In addition to the evident larger power, the Kolmogorov Smirnov approach has the further benefit of indicating the direction of the inhomogeneity. As we have seen, for species *Scholtzia involucrata* it indicates inhomogeneity in y-direction and for species *Conostephium pendulum* it indicates inhomogeneity in x-direction. For species *Andersonia heterophylla* it reveals inhomogeneity

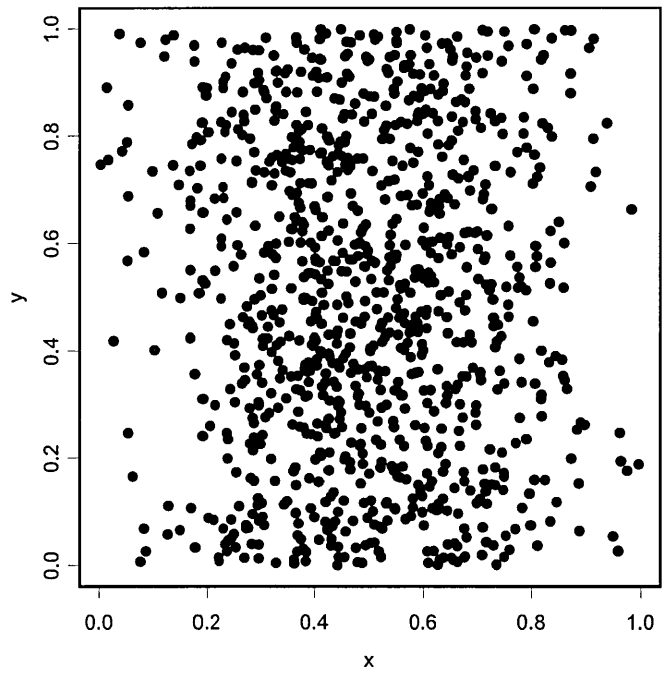

Figure 4.8: Simulated pattern with Gaussian trend in x-direction ($\mu = 0.5, \sigma = 2$)

in both directions. Based on this consideration it may be suggested that the classical F-test may be applied in both directions, as done in Section 4.1.2.3. Nevertheless, the approach does not test for more general trends or inhomogeneity. Whereas a north-west to south-east trend may be revealed as a trend in x- and y-direction, inhomogeneity along a diagonal line in the plot may not be detected.

In the following analyses we have treated all patterns for which the Kolmogorov Smirnov or the permutation test indicated inhomogeneity as in-

homogeneous. There was no pattern for which the F-test was significant and the other two tests were not. In addition, we have inspected the estimated intensity surfaces to ensure that all inhomogeneous patterns have been detected but found no further patterns which the tests may not have detected. Overall, out of the 35 patterns considered here, 23 were inhomogeneous according to this criterion. Two patterns showed inhomogeneity in x-direction only, 13 in y-direction only and five in both directions[9]. In Section 4.3, the results from the first and second order summary statistics for all species with at least 20 individuals are summarised and they are discussed from an ecological perspective in Section 4.5.3.

4.2 Second order summary statistics

Informed by the results from Section 4.1.2.3 the homogeneous or the inhomogeneous $L-$function as well as the pair correlation function were calculated for each of the subpatterns. More specifically, for those patterns where the Kolmogorov Smirnov test and/ or the permutation test indicated inhomogeneity, only the inhomogeneous L-function was calculated as the homogeneous version may indicate spurious clustering. For those patterns for which the tests had not indicated inhomogeneity, both homogeneous and inhomogeneous L-functions were estimated. This is because despite the insignificant test for inhomogeneity we cannot prove that the patterns are indeed homogeneous. The theoretical inhomogeneous L-function is a generalisation of the homogeneous L-function and should yield the same result as its homogeneous

[9]As there were three patterns for which only the permutation test was significant we do not know the direction of this inhomogeneity.

version in the special case of a homogeneous pattern. However, the estimation of the inhomogeneous L-function involves the estimation of the intensity surface which in turn requires a choice of a bandwidth (see Section 3.3.3.2), making the estimation procedure more susceptible to errors. We should thus expect the results to vary slightly. The pair correlation function is valid both for the inhomogeneous and the homogeneous case so no distinction has to be made.

Due to the large number of subpatterns, we present only a small number of the plots in this section. See Appendix B.2 for the estimated functions for those subprocesses considered in the thesis. For all cases we present the summary statistics along with a set of simulation envelopes derived from 39 simulations of homogeneous and inhomogeneous Poisson processes, respectively, with the same intensity as the observed pattern. See Section 3.6 for details on the construction of these envelopes.

4.2.1 Examples of second order summary statistics of homogeneous patterns

Figure 4.9 shows a plot of the homogeneous L-function and Figure 4.10 a plot of the inhomogeneous L-function and the pair correlation function for the species *Hibbertia hypericoides*. In all three plots the full line represents the estimated function for the species, the dashed lines represent the maximum, mean and minimum of the 39 simulated patterns at each distance, respectively. The species shows a random pattern, as both the homogeneous and the inhomogeneous L-function as well as the pair correlation function remain inside the simulation envelopes for all distances. We also observe

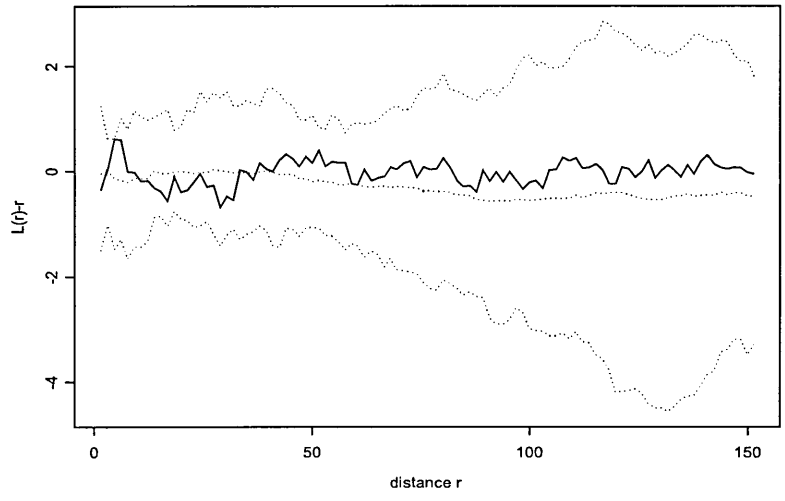

Figure 4.9: Homogeneous L-function with simulation envelopes for *Hibbertia hypericoides*, where r denotes distance in units of 10cm.

that the envelopes for both L-functions become wider for larger distances, indicating an increasing bias at larger distances, in particular for the inhomogeneous L-function. This is due to the smaller number of points taken into account at larger distances. This bias is absent from the plot of the pair correlation function. This difference between the two second order summary statistics will be discussed further and become particularly relevant in the analysis done in Chapter 5.

Inspecting the same type of plots for the species *Conostylis canescens*, the pattern clearly shows clustering; see Figure 4.11 for a plot of the homogeneous L-function and Figure 4.12 for a plot of the inhomogeneous L-function and the pair correlation function. All three plots indicate that the pattern shows clustering at close distances as the functions are above the simulation

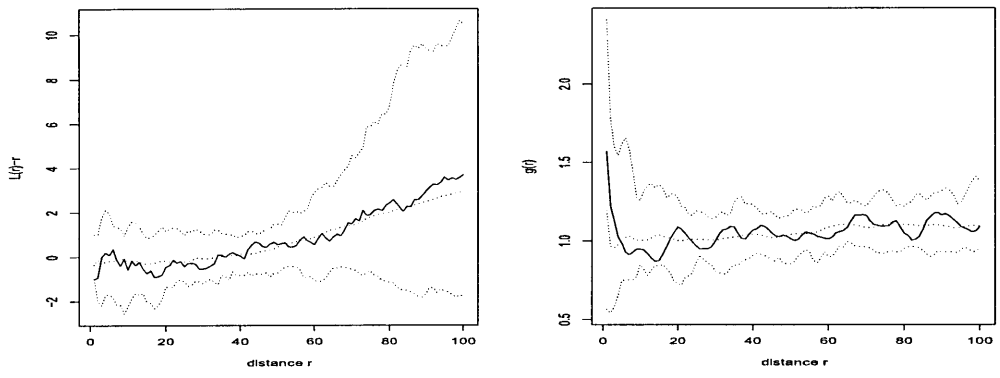

Figure 4.10: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Hibbertia hypericoides*, where r denotes distance in units of 10cm.

envelopes in all three cases. Distance is plotted here in units of 10cm. Due to the cumulative nature of the L-functions, these are above the envelopes up to a distance of about 16 units, i.e. 1.6 metres. The pair correlation function, on the other hand, remains above the envelopes up to a distance of 60cm only.

Similar plots were produced for all homogeneous species, with an abundance greater than 20. None of the species in the pattern showed a regular and homogeneous pattern. In total, seven subpatterns were classified as random and homogeneous and four as clustered and homogeneous (Appendix B.2). In Section 4.3, we summarise the results for all species and in Section 4.5.3. We discuss these results from an ecological perspective and relate these observations to properties of the respective species.

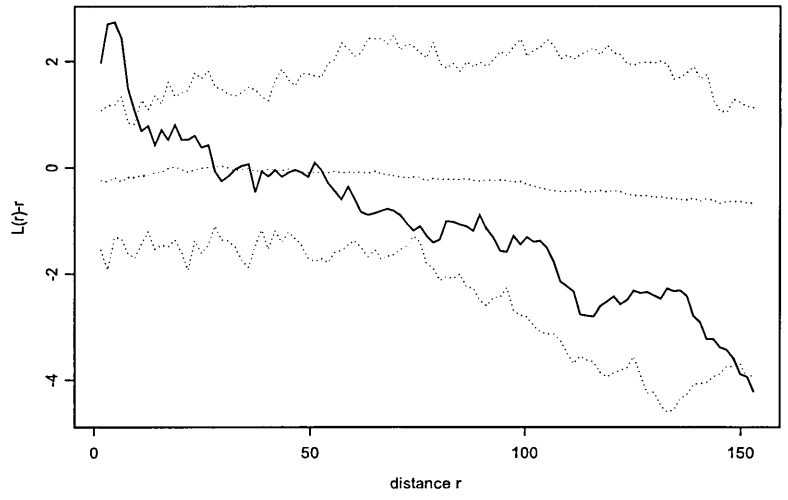

Figure 4.11: Homogeneous L-function with simulation envelopes for *Conostylis canescens*, where r denotes distance in units of 10cm.

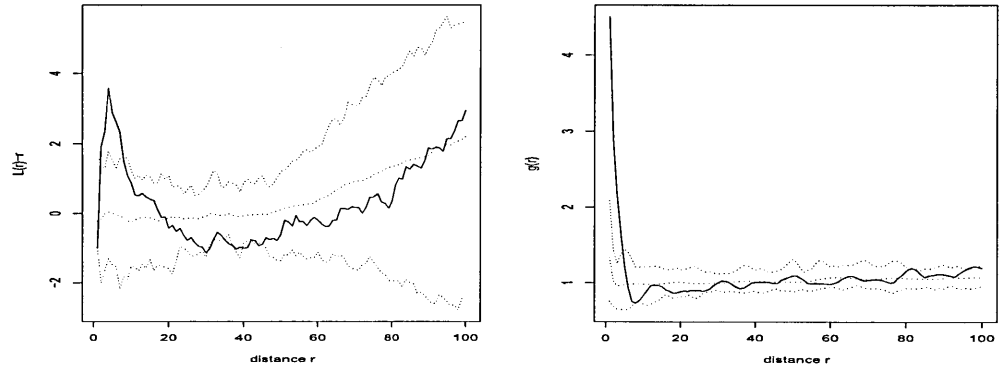

Figure 4.12: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Conostylis canescens*, where r denotes distance in units of 10cm.

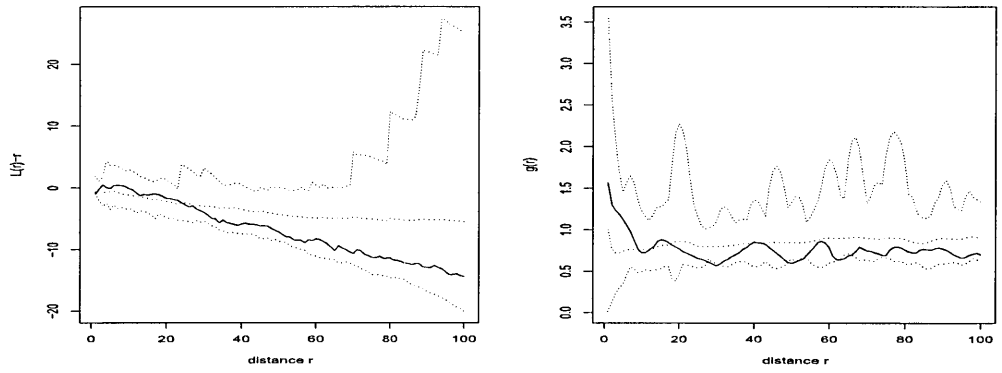

Figure 4.13: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Astroloma xerophyllum*, where r denotes distance in units of 10cm

4.2.2 Examples of second order summary statistics of inhomogeneous patterns

The L-function and the pair correlation function for species *Astroloma xerophyllum* are consistent with a random pattern, see Figure 4.13. Both functions remain within the envelopes at all distances, i.e. taking into account the results in Table 4.1, *Astroloma xerophyllum* can be classified as an inhomogeneous random pattern. We observe again the apparent bias in the estimation of the L-function, indicated by the wider envelopes at larger distances.

However, the species *Dasypogon bromeliifolius* shows a **clustered pattern**, see Figure 4.14 for a plot of the inhomogeneous L-function and the pair correlation function. Both summary statistics indicate that there are more plants at close distances than would have been expected from a random pattern. The pair-correlation function is above the envelopes up to a distance of 1.7 metres. The L-function shows clustering up to a distance of 3.6 metres.

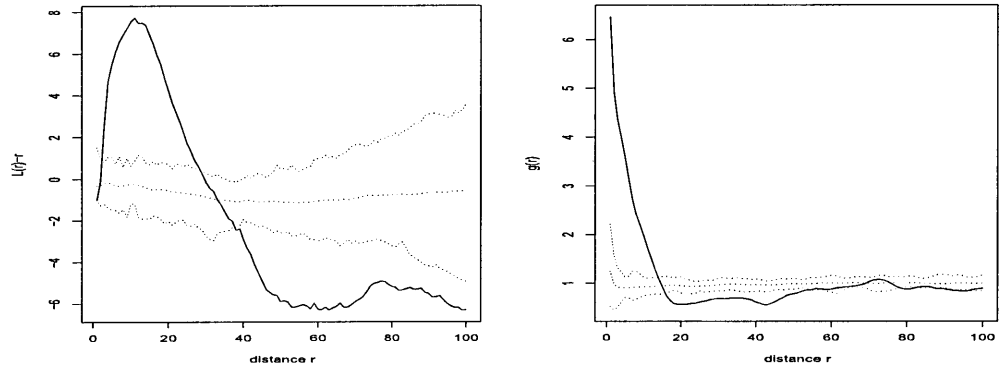

Figure 4.14: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Dasypogon bromeliifolius*, where r denotes distance in units of 10cm

This apparent difference is due to the cumulative nature of the L-function, which might result in misleading interpretations.

None of the species in the pattern showed a repulsive and inhomogeneous pattern. There were altogether 13 random inhomogeneous subpatterns and 12 clustered inhomogeneous patterns (Appendix B.2). As mentioned above, we summarise the results for all species in Section 4.3. In Section 4.5.3, we discuss these results from an ecological perspective and relate these observations to properties of the respective species.

We have now considered the first and second order properties of all species whose abundance was large enough to be considered interpretable. This process has been rather time-consuming and has required us to inspect a large number of functions. In the Section 4.4 we will devise a mechanism which enables us to detect quickly those patterns within a large data set which are "interesting", i.e. either inhomogeneous or non-random.

4.3 Summary of results of initial exploratory data analysis

This section summarises the results the above methods yield for all species[10]. Table 4.3 summarises the spatial behaviour of all species, and groups them by the characteristics of their pattern, distinguishing between homogeneous and inhomogeneous as well as random and clustered patterns. Note that none of the methods applied above identified any of the patterns as exhibiting a regular pattern.

[10] As above we have excluded species with less than 20 points from the analysis as these are considered too sparse to justify an interpretable analysis of the pattern.

	homogeneous	inhomogeneous
random	*Banksia attenuata*	*Alexgeogea nitens*
	Banksia menziesii	*Astroloma xerophyllum*
	Bossia eriocarpa	*Conostephium pendulum*
	Conospermum crassinervium	*Conostylis juncea*
	Hibbertia hypericoides	*Conostephium pendulum*
	Isopogon linearis	*Dampiera linearis*
	Lepidosperma tenue	*Hypocalymma xanthopetalum*
		Jacksonia floribunda
		Lepidosperma angistatum
		Leucopogon conostephioides
		Lomandra sp.
		Platysace juncea
		Chordifex sinuosus
		Scholtzia involucrata
clustered	*Conostylis canesence*	*Adenatos cygnorum*
	Hibbertia crassifolia	*Andersonia heterophylla*
	Phebocarya filifolia	*Boronia ramosa*
	Stylidium crossocephalum	*Dasypogon bromeliifolius*
		Eremaea astrocarpa
		Eriostemon spicatus
		Hibbertia sp.
		Leucopogon striatus
		Lyginia barbata
		Melaleuca scabra
		Patersonia occidentalis

Table 4.2: Summary of results from first and second order summary statistics

4.4 The CUSUM method for L-functions

In view of the large number of species in the data set and other data sets of even larger size (Burslem et al. 2001) an efficient method is sought that may quickly find those species whose spatial pattern is worth more detailed study. We aim to devise a method that may be used as an initial diagnostic

tool to filter out "interesting" patterns; in general, those patterns whose L-function differs from the L-function of a process exhibiting complete spatial randomness are interpreted as interesting in the given context[11].

4.4.1 Approach

CUSUM charts are a tool used in automated statistical process control where the deviation from a target is monitored (Wetherill and Brown 1991), e.g. in a production process where items have to be produced according to a specified size.

In this novel use of CUSUM, we interpret the L-function for a homogeneous Poisson process, i.e. the constant function $f(r) = r$, where r is the distance, as our target and cumulate (upper and lower) deviations of the estimated L-function $\hat{L}(r)$ from $L(r) = r$ to detect non-random patterns.

Note that this approach is different from the traditional CUSUM approach, as we consider a fixed value at each distance as the target, and not a parameter of a specific distribution. Furthermore, in most classical applications of the CUSUM method the target parameter is chosen such that properties of the distribution of the CUSUM statistic are known. This is not true in the case here. Since the distribution of the L-function is not known, the distribution of the CUSUM statistic is not known either. As a result, we use permutation methods instead to simulate the distribution of any test statistics in connection with the CUSUM charts.

[11] A similar approach could have been applied to the K-function and the pair-correlation function but we restrict the following approach to the L-function as we expect that the L-function is more sensitive than the pair correlation function due to its cumulative nature and has a more stable variance than the K-function.

More specifically, for distances r_i, with $r_1 < r_2 < r_3 < \ldots r_{n_{dist}}$, $i = 1, \ldots, n_{dist}$ we compare the estimator of the L-function $\hat{L}(r_i)$ of the respective process to the identity function: $z(r_i) = \hat{L}(r_i) - r_i$ and define the **upper CUSUM** $UCU(r_i)$ and the **lower CUSUM** $LCU(r_i)$ at distance r_i as

$$\begin{aligned} UCU(r_i) &= \max(0, z(r_i) + UCU(r_{i-1})) \\ LCU(r_i) &= \max(0, -z(r_i) + LCU(r_{i-1})). \end{aligned}$$

These are then used to define the following test statistics:

$$M_{UCU} = \max_i(UCU(r_i)) \tag{4.1}$$

and

$$M_{LCU} = \max_i(LCU(r_i)). \tag{4.2}$$

As mentioned above, the theoretical distributions for these test statistics are complex and not known explicitly since the distribution of the L-function is not known explicitly (Stoyan et al. 1995). For this reason the distributions are determined by simulation, using the following steps:

1. Simulate $m = 1000$ homogeneous Poisson processes;
2. calculate M_{UCU} and M_{LCU} for these to yield the empirical distributions for M_{UCU} and M_{LCU} for these;
3. determine critical values at the chosen α-level[12].

[12]The choice of the alpha level will depend on whether the method is applied in a confirmatory or in an exploratory setting; given the exploratory nature of the approach here, an alpha level of $\alpha > 0.05$ might be considered appropriate.

Note that the distributions for M_{UCU} and M_{LCU} are different and thus have to be determined separately. This is due to the fact that the L-function is bounded by $-r$ below the theoretical line at each distance r but the deviation above the theoretical line is not bounded. As a result, estimated L-functions for clustered patterns will deviate more from the "target" than those for regular patterns. Section 5.1.3 will discuss this issue in more detail as it becomes particularly relevant in the methodology developed there.

Also note that this approach is not meant to distinguish explicitly and simultaneously between clustered, random and regular patterns and detect whether these are inhomogeneous or homogeneous. Whereas a significant lower CUSUM will indicate regularity, a significant upper CUSUM does not necessarily indicate clustering but may merely indicate spurious clustering resulting from inhomogeneity. We will recommend the use of the approach as an initial exploratory data analysis tool irrespective of the homogeneity or inhomogeneity of the pattern. This can be done since the homogeneous L-function applied to an inhomogeneous pattern will still be different from the theoretical L-function for a homogeneous Poisson process. Figure 4.15 illustrates this point. Here, the homogeneous L-function has been applied to a pattern simulated from an inhomogeneous Poisson process (see Figure 4.15 (a) for the pattern and 4.15 (b) for the estimated L-function). The L-function is outside and above the envelopes for a large range of distances indicating clustering, which is spurious as the pattern has been generated from a Poisson process. Even though this is a misleading result the CUSUM approach when applied to this pattern will indicate that this is an "interesting" pattern that needs to be studied in more detail. Thus the method may detect the

deviation and suggest further inspection of the relevant pattern.

The CUSUM method may be applied to the inhomogeneous L-functions, serving as a formal test of second order summary statistics. However the benefits of the CUSUM approach as a means to detect quickly patterns of interest will be lost in this case, as the appropriate estimation of inhomogeneous L-functions requires the careful choice of a bandwidth which involves the careful inspection of estimated intensity surfaces with different bandwidths making the procedure much slower.

4.4.2 Statistical power – simulation study

A power study was conducted in order to determine the method's performance, in particular in comparison with the **max-dist test** T, the standard method for testing for non-randomness described in Section 3.3.5 above, which looks at the maximum (absolute) distance of the estimated L-function from the theoretical value expected for a homogeneous Poisson process.

Figure 4.16 show the results of the power study for patterns generated from a Strauss process with varying degree of repulsion reflected in the parameter γ. Note that for $\gamma = 1$ there is no repulsion and for $\gamma = 0$ there is complete repulsion with the radius r; see Section 3.4.4.1 above for more details on the Strauss process.

Figure 4.16 a) shows the results based on generated patterns with $n = 100$ points and Figure 4.16 b) shows the results based on generated patterns with $n = 200$ points. The full line shows the percentage of regular pattern detected by the original method T and the dashed line the percentage detected by the CUSUM method.

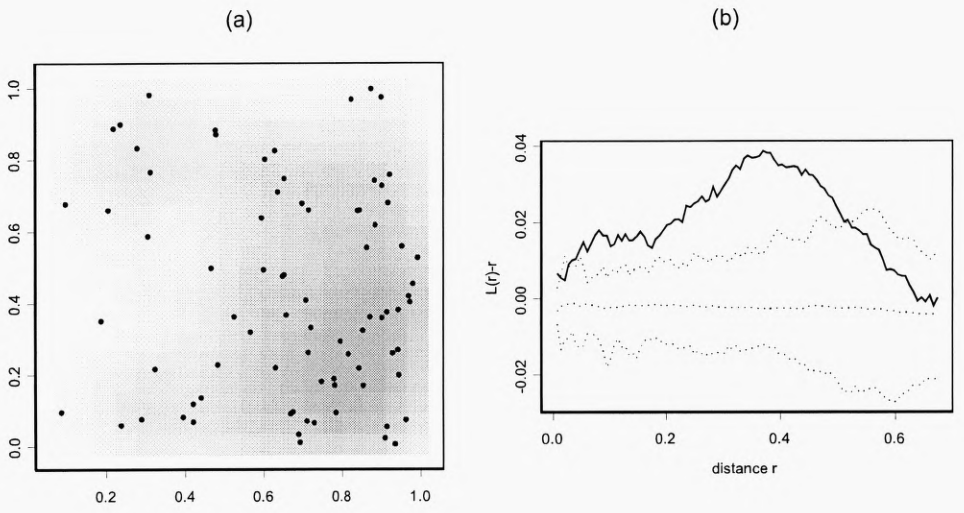

Figure 4.15: Pattern generated from an inhomogeneous Poisson process with estimated intensity surface (a) and homogeneous L-function for this pattern (b).

It is clear from Figure 4.16 that the CUSUM method identifies a higher proportion of patterns as non-random than the max-dist test when γ is closer to one. This indicates that it has greater statistical power than the standard method for low degrees of repulsion (large values of γ), i.e. when the patterns tend to be more similar to a random pattern. For patterns with stronger regularity we get the opposite effect with greater statistical power for the standard method than for the CUSUM approach, most notably for lower values of n. This result indicates that the CUSUM method is particularly

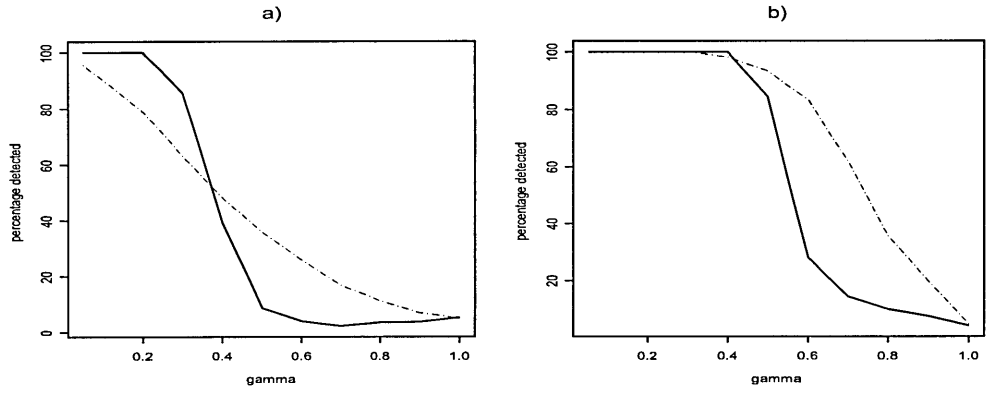

Figure 4.16: Results from power study, using simulated Strauss processes with varying degree of repulsion for patterns with n =100 (a) and n = 200 (b) points. The full line refers to the original max-dist test and the dashed line to the CUSUM approach; large values of γ indicate weak regularity.

useful when applied to patterns that are very similar to random patterns, i.e. those which cannot easily be detected by visual inspection of plots of estimated L-functions. A reason for this phenomenon may be the cumulative nature of the CUSUM approach which emphasises repeated deviations of the L-function from the "target" line, the theoretical L-function, in the same direction. In the Cooljarloo data set no pattern was identified as regular so this result from the simulation study is not directly relevant here but may be useful in the context of other data sets.

A similar power study was performed for clustered patterns. Figure 4.17 shows the results of the power study for 20 patterns generated from a Thomas process with weak and strong clustering, respectively. Here, the mean num-

ber of points per cluster was kept constant such that strength of clustering was interpreted as equivalent to size of clusters.

Figure 4.17 a) shows the results based on generated patterns with $n = 100$ points and Figure 4.17 b) shows the results based on generated patterns with $n = 200$ points. The full line shows the percentage of clustered patterns identified by the original method and the dashed line the percentage identified by the CUSUM method.

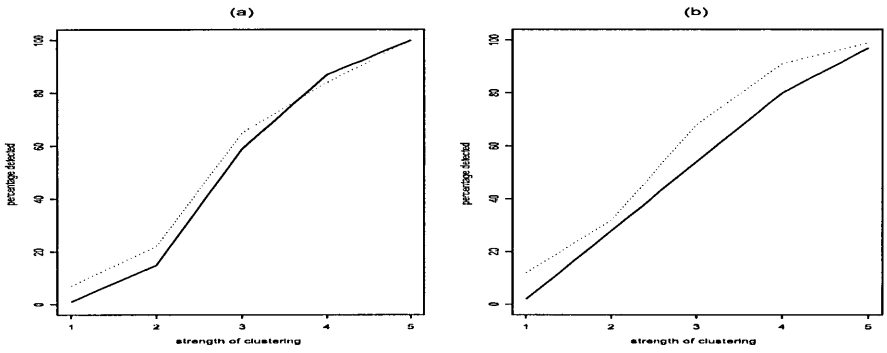

Figure 4.17: Results from power study, using simulated Thomas processes with varying degree of clustering with n =100 (a) and n = 200 (b) points. The full line refers to the original max-dist test and the dashed line to the CUSUM approach; large values on the y-axis indicate strong clustering.

Similar to the situation with regular patterns in Figure 4.17, the CUSUM method identifies a higher proportion of patterns as clustered than the original method when the patterns are less clustered. In addition, it also typically identifies more patterns as clustered, when the clustering is more pronounced. This indicates that the method may have greater statistical power than the standard method when used to identify clustered patterns.

4.4.3 Results with dataset

The CUSUM procedure described above was applied to the Cooljarloo data set to identify sub-patterns that show an apparent deviation from a homogeneous Poisson process. Due to the explorative nature of the application, we aimed at using a non-conservative decision criterion and thus chose a significance level of $\alpha = 0.1$ as opposed to the more common $\alpha = 0.05$. Since we have a multiple testing problem similar to the one described in Section 4.1.2.2 we adjust the significance level accordingly. This results in a p-value being considered significant if either one or both of the p-values for the upper and lower CUSUM are smaller than 0.0513. Table 4.4.3 lists the results for selected species. As above, these include examples showing typical features of both the methods and the data set, as well as those species used in the modelling process described in Chapter 6. Tables B.3 and B.4 in Appendix B.2.2 summarise the results for all species with at least 20 plants.

When comparing the results for two methods we find that there are two cases where both methods detect deviations from complete spatial randomness. For the species *Banksia attenuata* and the species *Leucopogon conostephioides* both methods agree that these are consistent with homogeneous random patterns. For the *Banksia* this is in accordance with earlier results whereas previous results indicated that the latter species exhibits an inhomogeneous pattern (see Table 4.1). Whether this is due to a lack in power of the methods discussed in this section or an over-sensitivity of the earlier methods cannot be decided here. The methods also agree that the patterns formed by the species *Andersonia heterophylla*, *Jacksonia floribunda* and *Leucopogon*

species	max-dist test	upper CUSUM	lower CUSUM
Andersonia heterophylla	**p = 0.000**	**p = 0.000**	$p = 1$
Banksia attenuata	$p = 0.34$	$p = 0.20$	$p = 0.77$
Bossiaea eriocarpa	$p = 0.25$	**p = 0.01**	$p = 0.96$
Conospermum crassinervium	$p = 0.12$	**p = 0.000**	$p = 0.78$
Conostylis canescens	**p = 0.04**	**p = 0.052**	$p = 0.62$
Jacksonia floribunda	**p = 0.000**	**p = 0.000**	$p = 1$
Leucopogon conostephioides	$p = 0.81$	$p = 0.17$	$p = 0.76$
Leucopogon striatus	**p = 0.000**	**p = 0.000**	$p = 1$

Table 4.3: Comparison of results from the max-dist test and the CUSUM approach for selected species; significant p-values are in bold

striatus deviate from complete spatial randomness. All three species have previously been classified as showing inhomogeneity, *Andersonia heterophylla* being strongly clustered, *Leucopogon striatus* weakly clustered and *Jacksonia floribunda* random. There is a number of species for which the CUSUM method yields significant results but the max-dist test does not, e.g. *Bossiaea eriocarpa* and *Conospermum crassinervium*. According to the methods applied in the sections above, both species exhibit inhomogeneous random patterns. Hence, the test has been sucessfully applied to detect inhomogeneity.

In consistence with the results from the application of second order summary

statistics with simulation envelopes, the CUSUM method indicated regularity for none of the patterns. Also, there was only one case (the species *Lepidosperma tenue*) which had been considered to show complete spatial randomness before but has been identified as not showing complete spatial randomness by the CUSUM method.

4.4.4 Two dimensional L-functions and pair correlation functions

Theoretically, both two-dimensional L-functions and pair correlation functions, as introduced in 3.3.4, may be estimated for all pairs of species with at least 20 individuals, but this would result in 561 pairs of plots.[13] We have thus chosen a few pairs of species that have been suspected to interact and will display the pair correlation functions only. More specifically, we focused on the resprouting species *Scholtzia involucrata* (species number 61), which was assumed to repulse most other species as it has been established in the same location for a very long time and hence a well-developed root system (Armstrong 1991). Note that the method is not capable of determining the direction of a potential repulsion or attraction; it is not clear whether one species attracts/repulses the other or *vice versa*. See Figure 4.18 for the pair-correlation functions for *Scholtzia involucrata* and the seeder species *Leucopogon striatus* (species number 48). Similarly, Figure 4.19 shows the same for *Scholtzia involucrata* and resprouting species *Lyginia barbata* (species number 50).

[13]Note that the estimated two-dimensional second order summary statistics $\hat{L}_{ij}$ and $\hat{L}_{ji}$ may differ slightly as a result of estimation bias whereas for the theoretical functions we have $L_{ij} = L_{ji}$ (Møller and Waagepetersen 2003b).

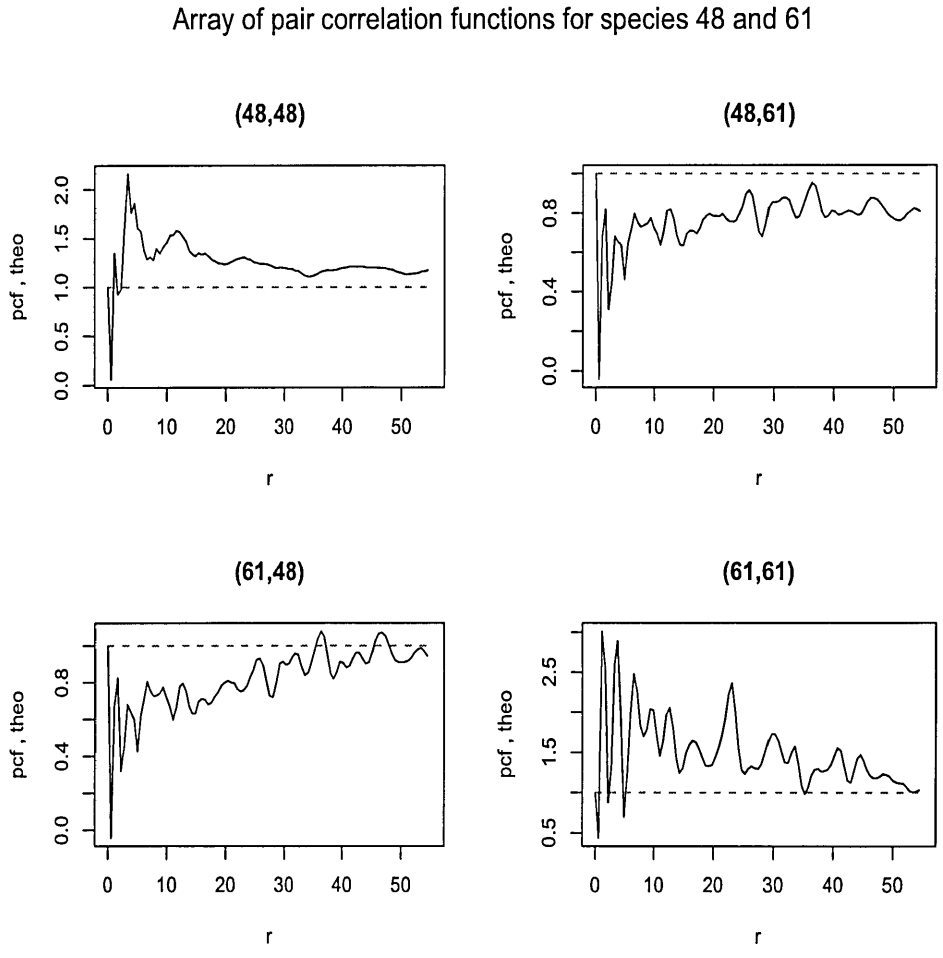

Figure 4.18: Pair correlation function for *Scholtzia involucrata* (number 61) and *Leucopogon striatus* (number 48), distance r in 10cm

It is clear from Figure 4.18 that there is repulsion between the species *Scholtzia involucrata* and *Leucopogon striatus* as the estimated two-dimensional pair correlation functions are below 1 for all distances considered. Fi-

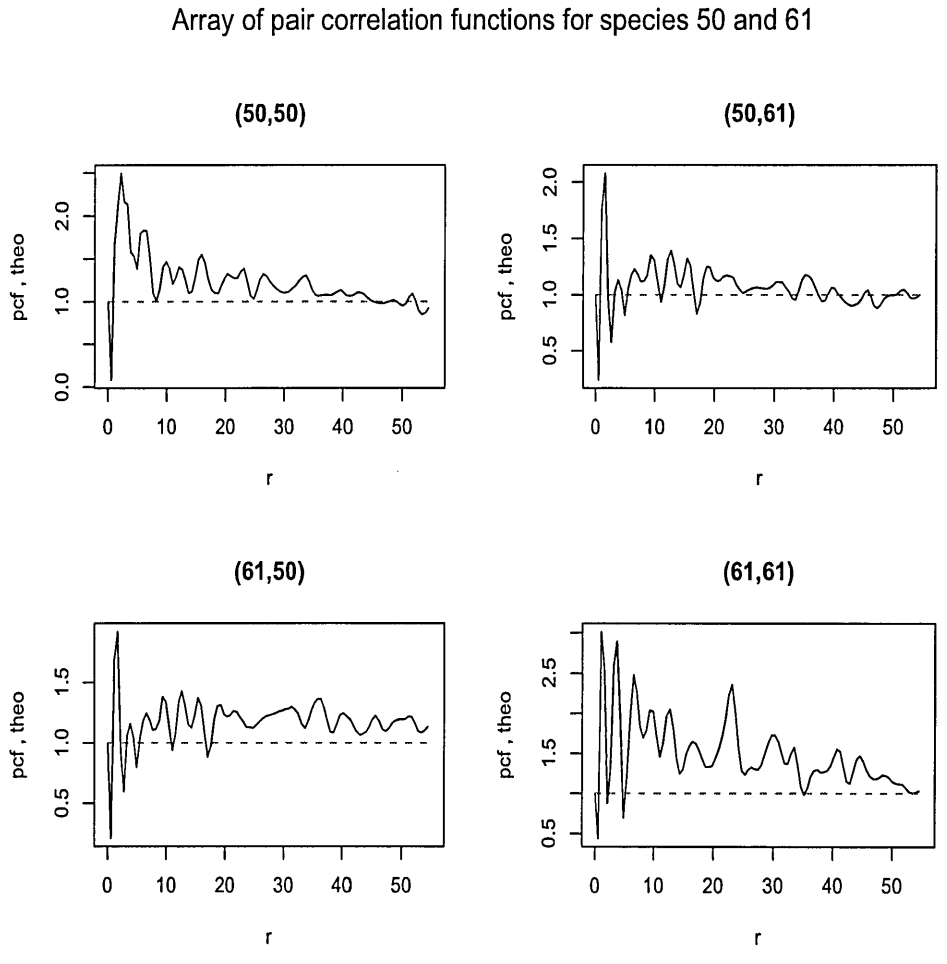

Figure 4.19: Pair correlation function for *Scholtzia involucrata* (number 61) and *Lyginia barbata* (number 50), distance r in 10cm

gure 4.19, on the other hand reveals, that the species *Scholtzia involucrata* (number 61) and *Lyginia barbata* do not exhibit mutual interaction as the estimated function hovers around the value of 1 for all distances considered. The models in Chapter 6 look further into inter-species interaction; the approaches applied in Sections 6.2 and 6.3 consider interactions between pairs

of species while still taking the other species into account.

4.5 Discussion

This chapter presented the exploratory data analysis of the Cooljarloo data set, establishing an overview of the first and second order characteristics of the subpatterns formed by the species. We close this chapter with some comments on the methodology itself (Section 4.5.1), a discussion of the results in the light of ecological theory (Section 4.5.2) and a discussion of the results with the data set (Section 4.5.3).

4.5.1 Discussion of methodology

As far as first order summary statistics are concerned, the results in Section 4.1.2.3 clearly show that traditional tests for inhomogenity of spatial point patterns may not be sensitive enough and alternative approaches need to be sought. The methods described here, most notably the Kolmogorov Smirnov approach, have shown promising results, i.e. seemed to be more powerful in detecting inhomogeneity in the form of a linear trend. However, more detailed power studies would have to be done to determine thoroughly the properties of the different testing approaches and related methods including other goodness of fit test. Further, other types of alternatives such as Gaussian trend rather than linear trend would have to be taken into account.

In the context of second order summary statistics, the CUSUM method turned out to be a very flexible tool. The permutation test necessary for estimating the distribution of the upper and lower CUSUM requires simula-

tions. These might take some time to run, in particular when the number of points is large. However, the procedure can be automated so does not require any manual interference or careful visual inspection. Of course, the approach here is only the first step in a more thorough analysis but it has turned out to be parsimonious and thus particularly useful in the current context, i.e. when dealing with a highly multivariate data set. A natural extension of the method would be to apply it to pair-wise second order summary statistics, to filter out pair-wise interaction between species which seem to be of interest. Similarly, other second order summary statistics may be used instead of the L-function and further improve the power of the test.

4.5.2 Discussion of biodiversity theory in this context

The analysis of the estimated L-functions and pair correlation functions revealed a variation of intra-specific as well as inter-specific interaction among species. Neutral theory postulates that the identity of an individual does not influence the interaction structure and hence community dynamics, whereas the niche theory claims that it makes a difference which other species a certain species interacts with and hence that speciation has an influence on community dynamics. However, our findings indicate that interaction strength and direction varies with the identity of the involved species with negative, positive and no interaction within a species and more importantly between pairs of species. All these aspects yield arguments in favour of niche theory for this data set.

Further, the analysis of the estimated L-functions and pair correlation functions indicated that a large number of species exhibit spatial clustering.

The Janzen-Connell hypothesis predicts that species' survival is improved with increasing distance from the location of a conspecific adult tree and decreases with the number of conspecific individuals in close vicinity. This would predict that older individuals would tend to form a more regular pattern than juvenile individuals of the same species which might be clustered as a result of initial seed dispersal.

The age structure in the given data set is typical in Western Australia but is very different from age structures observed in other natural systems. The resprouting species are generally extremely old and the individuals from seeder species extremely young, not older than the time since the last fire, i.e. ten years in the Cooljarloo data set. Furthermore, individuals from the same seeder species tend to be of the same age as their germination is directly linked to the occurrence of the fire. As a result, we cannot compare the patterning of juvenile versus older individuals, even if data on plant age were available. A comparison of clustering strength between resprouters and seeders is not meaningful since growth behaviour, which strongly influences spatial patterning, varies strongly among the species rendering a direct comparison infeasible. Nevertheless, the fact that a number of resprouters, after having survived for extremely long times, still exhibit spatial clustering may provide evidence against the Janzen-Connell hypothesis (Condit et al. 2000; Hubbell 1979).

4.5.3 Discussion of results for data set

Most of the species in the data set showed an inhomogeneous pattern. The soil can be considered homogeneous with regard to nutrient levels (Dixon

2005). Hence, species' inhomogeneity is very unlikely to be a result of inhomogeneous nutrient levels but rather a result of species' growth habits and reproduction patterns or intra-species interaction, such as repulsion or attraction between different species.

The considered species are known to occur in large-scale patches such that large clusters of species are typical (Armstrong 2005; Dixon 2005). Thus, in most cases where an inhomogeneity has been observed, it is unlikely to be due to a gradient in soil nutrients but should be considered as large scale clustering where only part of a bigger cluster has been observed in the plot.

For those species which were found to be homogeneous the plot analysed here is likely to have been positioned within a large cluster in which no gradient exists. For instance, the species *Banksia attenuata* is a tree species and operates on a larger scale than most of the other shrub species which in turn makes it less likely that we have observed the "edge" of a cluster within this plot.

The second order summary statistics have been corrected for inhomogeneity, which here can be translated into large-scale clustering. Any perceived clustering found from the second order summary statistics indicates smaller scale clustering which results from reproduction patterns, such as seed dispersal and growth habits. All these results should impact on re-naturation practices. Both a species' tendency towards large- and small-scale clustering as well as repulsion or attraction between pairs of species should be taken into account in any re-naturation attempt.

Chapter 5 will look further into inter-species interactions, whereas in Chapter 6 we will mainly focus on analysing inter-species interaction by applying a hierarchical spatial point process model.

Chapter 5

Principal component analysis for spatial point patterns

In the previous chapter we have considered only the spatial pattern of individual species or pairs of species. Here, we now provide an overall characterisation of the intra-species interaction structure for the whole community. We reveal the most common types of pattern formation as well as those spatial characteristics that vary most among species. By grouping species according to these characteristics, we gain an understanding as to which aspects of spatial patterning are the most distinctive between species in the community. This may be indicative of a niche specification having developed with regard to the species' spatial behaviour (see Section 2.3.1).

The case study data set consists of the locations of a large number of species, i.e. it is highly multivariate, and consequently highly complex. We seek to address this complexity by reducing the dimensionality of the data set by grouping the species according to the spatial patterning formed by the individual species. Ideally, this grouping should be based on those aspects of the spatial formation that are most distinctive among the species.

Furthermore the grouping should be more refined than the rough grouping described in Table 4.2. It should yield a detailed similarity structure of the patterns within the same group, i.e. indicate the position of each of the species relative to the other members of the group with regard to similarity in spatial patterning. Further, the grouping should also reflect more refined characteristics of each species' spatial pattern, e.g. clustering at closer distances versus clustering at larger distances rather than just clustering. A refined knowledge of the spatial pattern in the community may also inform the further modelling process as species with a similar spatial behaviour may be modelled with a similar spatial point process model[1].

The concept and approach of functional principal component analysis and its application to second order summary statistics is outlined (5.1). The capability of the approach to characterise and group known patterns is assessed prior to its application to the case study data set (5.2). Further, the robustness to noise in the data is assessed (5.3), Illian et al. (2004). Finally, the approach is applied to the case study data set (5.4) and the results discussed in terms of the approach's descriptive capability (5.5), Illian et al. 2005.

5.1 Methods

In traditional statistics, multivariate approaches such as principal component analysis (PCA) are used for dimension reduction (Jolliffe 2002). Similar me-

[1]Note that in this chapter, we only include those patterns with number of points ≥ 25, since we consider patterns with a smaller number of points too sparse to yield interpretable information on their spatial formation. In the given data set this excluded 36 species out of 67 from the formal analysis. As mentioned above we shall see in Chapter 6 that sparser species may still be included in a model as they may influence the pattern of other less sparse species.

thods do not exist for spatial point pattern data and thus are developed here for the first time, using methods from functional data analysis on the second order summary statistics. The following section describes the basic ideas of functional data analysis in general (5.1.1) and Section 5.1.2 introduces functional principal component analysis (FPCA). Section 5.1.3 explains how this may be applied to second order summary statistics.

5.1.1 Functional data analysis

We only introduce the basic ideas of functional data analysis; for a more detailed introduction to functional data analysis see Ramsay and Silverman (1997, 2002). Functional data analysis operates on functional data, i.e. observations that are functions interpreted as single entities rather than as consecutive measurements. Generally, the record of a functional observation x consists of n pairs $(t_j, y_j), j = 1, \ldots, n$, where y_j is an observation of the function $x(t_j)$ at time t_j. Since the functions are usually observed at a finite number of values of t_j only, interpolation or smoothing techniques have to be applied to yield a functional representation of the data. A standard approach here is the basis function method where a function is represented by a linear combination of K unknown basis functions ψ_k, i.e.

$$x(t) = \sum_{k=1}^{K} c_k \psi_k(t).$$

The coefficients c_k of the expansion are determined by minimising the least square criterion

$$LS(y|\mathbf{c}) = \sum_{j=1}^{n} [y_j - \sum_{k=1}^{K} c_k \psi_k(t_j)]^2$$

or in matrix form:

$$LS(y|\mathbf{c}) = \|y - \Psi\mathbf{c}\|^2,$$

where the vector $\mathbf{c}$ contains the coefficients c_k and $\Psi = \{\psi_k(t_j)\}$. LS can be minimised by the solution $\mathbf{c} = (\Psi'\Psi)^{-1}\Psi'y$. In general, the degree of smoothing depends on the number of basis functions. There is a wide range of basis functions to choose from, e.g. Fourier series, polynomial bases, splines, wavelets etc., see Ramsay and Silverman (1997) for more details. A number of standard statistical methods have been generalised to be applied to functional data, e.g. analysis of variance, discriminant analysis and principal component analysis. We will apply the latter for our purposes, as it is best suited for both dimension reduction and revealing in which aspect of their spatial behaviour the species vary the most.

5.1.2 Functional principal component analysis

In this section, we recall the concept and aims of principal component analysis in a non-functional setting in order to describe its purpose and generalisation to a functional context. Principal component analysis in a non-functional setting is a multivariate method that identifies a lower-dimensional representation of high-dimensional data by constructing a small number of linear combinations of the original large number of variables. Those combinations are chosen that the lower dimensional representation best explains the variance in the data (Jolliffe 2002).

More technically, we consider $\mathbf{u}_i = (u_{i1}, \ldots, u_{ip})$, the vector of replicates $i = 1, \ldots, N$ on variables $j = 1, \ldots, p$ and linear combinations $f_i = \sum_{j=1}^{p} \beta_j u_{ij}, \quad i = 1, \ldots, N$ with weight vector $\beta = (\beta_1, \ldots, \beta_p)$. We

need to find a weight vector $w_1 = (w_{11}, \ldots, w_{p1})$ such that $N^{-1} \sum_i f_{i1}^2(.)$ is maximal under the restriction $\|w_1\|^2 = 1$. In subsequent steps we repeat the above to find vectors $w_2, \ldots, w_m, \quad m \leq p$ with the new weight vectors being orthogonal to the previous ones. This is equivalent to finding the solution with largest eigenvalues λ to the eigenequation $V\mathbf{w} = \lambda\mathbf{w}$, where $V = N^{-1}U'U$ is the sample variance-covariance matrix, $U = u_{ij}$ and $\mathbf{w}$ a vector of weights.

For a functional PCA (see Ramsay and Silverman (1997, 2002), Jolliffe (2002)) the approach is very similar. Again linear combinations explaining as much variance as possible are sought. Since we are now in a continuous setting the sums are replaced by integrals and these "linear combinations" maximise the variance between the shape of the functions.

Technically, we thus consider function values $u_i(t)$ and define

$$f_i = \int \beta(t) u_i(t)\, dt,$$

where $\beta(t)$ is a weight function, and maximise $N^{-1} \sum_i^N f_{i1}^2$ under the constraint $\|w_1\|^2 = \int w_1(t)^2\, dt = 1$ and have an eigenequation

$$\int v(t,s) w(s)\, ds = \lambda w(t) \tag{5.1}$$

with variance-covariance function $v(t,s) = N^{-1} \sum_{i=1}^N u_i(t) u_i(s)$. The solution to this eigenequation with largest eigenvalue solves the maximisation problem. In the subsequent p steps maximise $N^{-1} \sum_i^N f_{ik}^2$, where $k = 1, \ldots, p$, subject to $\|w_i\|^2 = \int w_i(t)^2\, dt = 1$. Further analysis will examine the scores f_{ik} for each of the original smoothed curves on the first p principal compo-

nents, with $p \ll k$ in connection with an interpretation of the shape of these principal components.

5.1.3 Functional principal component analysis of second-order summary statistics

We now describe the method outlined in the previous section in the context of spatial point pattern data. This is done by applying FPCA to the second order summary statistics, i.e. the K- or L-function or the pair-correlation function. Since they are functions the estimated second order summary statistics of a multitype spatial point pattern can be considered functional data. Consequently, a functional principal component analysis on the summary statistics groups the statistics by their similarity in shape. Hence this yields a grouping of the subpatterns for the species in a multitype spatial point pattern by spatial behaviour with regard to second order characteristics.

More technically, let Z be a spatial point process on $\mathbb{R}^2$. Let X be a multitype point process $X = \{(\zeta, m_\zeta) : \zeta \in Z\}$ with $m_\zeta \in \mathcal{M}$ and $\mathcal{M} = \{1, \ldots, k\}$ a set, where no other marks are available, and subprocesses $X_i \subset X$ with $X_i = \{(\zeta, m_\zeta) : \zeta \in Z \text{ and } m_\zeta = i\}$ and $i = 1, \ldots, k$. Consider a realisation x of X. We use second-order summary statistics, in particular L-functions or pair correlation functions, to characterise the spatial behaviour of the individual subpatterns x_i. These statistics are estimated using the estimators described in Section 3.3.3.

We smooth the estimated second-order summary statistics using cubic B-splines; these were chosen as they are splines with compact support, capable

of picking up local features (see Green and Silverman (1994)). The splines may be used to approximate functions by fitting a set of polynomials with a degree of at most three between "knots" yielding a functional representation which has a continuous second derivative.
We subsequently perform a functional PCA on the smoothed functions. Through this, the subprocesses may be grouped on the basis of their scores on the principal components, where the number of principal components is determined by a scree plot (Jolliffe 2002).

We use hierarchical cluster analysis on these scores, in particular Ward's method (Everitt et al. 2001), to detect clusters of similar second-order summary statistics and hence groups of point processes with similar spatial behaviour. The result of the cluster analysis is plotted in a dendrogram and, together with a plot of the first p principal components, reveals groups of points processes with similar spatial behaviour. In addition, the finer structure of the dendrogram displays similarities between individual patterns within the groups. Thus, the dendrogram summarises the most distinctive features in the population as well as the position of the individual species with regard to these features relative to the other species.
Note that $L(r)$-values for regular patterns tend to lie in $[0, r]$ whereas $L(r)$-values for clustered patterns are usually larger than r with no upper bound. Hence, if we want to distinguish between clustered, random and regular patterns, the difference between the L-function for a clustered pattern and a random pattern tends to be larger than the difference between the L-function for a regular pattern and a random pattern. In analogy to the approach taken in a standard PCA context when variables have been measured on different

scales, we perform a FPCA on the correlation matrix rather than on the covariance matrix. I.e. equation (5.1) now becomes

$$\int v^*(t,s)w(s)\,ds = \lambda w(t),$$

where v^* is the correlation function $v(t,s) = N^{-1}\sum_{i=1}^{N} x_i^*(t)x_i^*(s)$, i.e. the covariance function of a standardised data matrix x^*.

A similar situation occurs when the pair-correlation function is being used – regular and random patterns appear more similar than clustered and random patterns so a FPCA on the correlation matrix will be used with both summary statistics. The simulation study in Section 5.2 has investigated this aspect and compares the performance of the two statistics in this context.

For illustration, Figure 5.1 depicts an example using simulated data of 20 clustered and 20 random patterns (for more details on the simulations in this context see Section 5.2). Figure 5.1 a) shows the smoothed L- functions, 5.1 b) the first two principal components, 5.1 c) a plot of the scores of all patterns on the first two principal components and 5.1 d) a dendrogram of the scores. The first principal component (full line, Figure 5.1 b)) represents clustering behaviour at close distances and the second principal component (dashed line) summarizes the behaviour at larger distances. This indicates that the different subpatterns differ most with regard to absence or presence of clustering at close distances, as expected. The dendrogram in Figure 5.1 d) clearly shows two distinct groups- the clustered patterns and the random patterns.

Note that here, as well as in all other examples including the data set the first two principal components explained more than 85% (for this example:

87.4%) of the variance, with the third PC only contributing very little (for this example: 0.78%). We have thus used only two PCs in most of the examples and do not comment on this in each case.

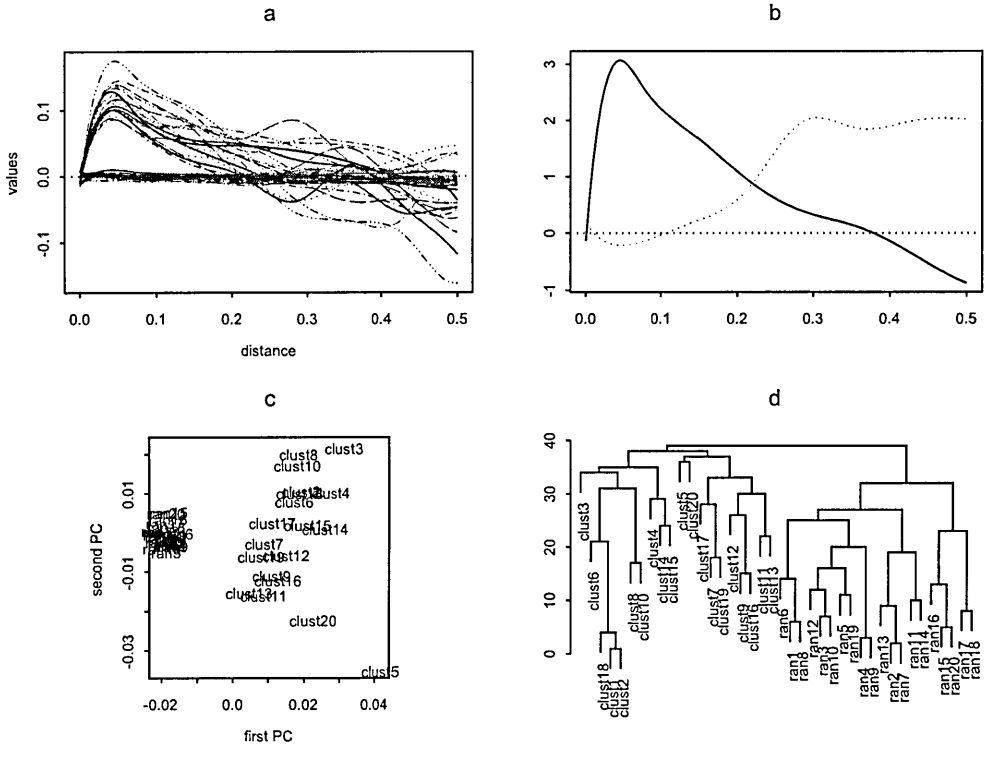

Figure 5.1: Results from a functional principal component analysis on L-functions for 20 clustered and 20 random patterns, a) smoothed L-functions, b) first two principal components, c) a plot of the scores of all patterns on the first two principal components, d) dendrogram of the scores

5.2 Feasibility study

Prior to application of the approach to the case study data set, we assess the ability of the approach to distinguish point patterns with known and different spatial behaviour. For this purpose we ran a feasibility study and simulated groups of point patterns with different but known properties. These included groups of random patterns, generated from homogeneous or inhomogeneous Poisson processes, clustered patterns, generated from Poisson cluster processes or Thomas processes, and regular patterns, generated from Strauss processes (refer to Section 3.4 above for more details on these processes). Overall, the following aspects are considered:

a) Before any functional data analysis, the estimated functions, i.e. the summary statistics here, must be smoothed as explained above (see Section 5.1.1). We use cubic B-splines for this purpose for reasons given in Section 5.1.3. Hence we need to advise on an appropriate choice for the number of splines used, as the number of splines determines the degree of smoothness.

b) Furthermore, the choice of the clustering algorithm applied to the scores on the principal components after the principal component analysis is crucial for the results as different algorithms yield different results.

c) It is not directly clear which of the two summary statistics, i.e. the L-function or the pair-correlation function, yield better results.

d) Finally, the capability of the overall method to distinguish groups of different spatial behaviour is to be investigated such that both its strengths and its weaknesses are revealed and recommendations made.

Aspects a) and b) are addressed by investigating the effect of the number of splines on the smoothness of the summary statistics and the effect of the clustering algorithm on the structure of the dendrograms respectively with various types of simulated patterns, see below.

Aspects c) and d), however, lead to a detailed and extended simulation study where the performance of the method was assessed both with homogeneous (Section 5.2.1.1) and inhomogeneous patterns (Section 5.2.1.2). The ability of the method to distinguish correctly between distinct groups of simulated patterns with known spatial behaviour is used as a criterion for the quality of its performance.

a) **Choice of number of splines** It is not possible to provide a general rule for determining the "right" number of splines for any application of spline smoothing (Green and Silverman 1994). The appropriate number will differ across applications and depends heavily on the number of points in which the function to be smoothed has been estimated. Generally speaking, a smaller number of splines yields very smooth functions which might ignore locally important features. However, using larger numbers of splines retains more detail but may also produce spurious oscillation.

Figure 5.2 shows an example where the L-functions for 20 simulated random patterns (generated from a Poisson process) and 20 simulated

regular patterns (generated from a Strauss process) with 100 points each were smoothed with 5, 8 and 20 splines respectively. The small number of splines in 5.2 a) leads to over-smoothing where the typical shape of the L-function and the fact that the function passes through the origin are lost, whereas 5.2 c) shows a case where too many splines have been used such that the functions oscillate too strongly. In this specific case the number of splines used in 5.2 b) may be the most suitable. Overall, the plots of the smoothed functions have to be carefully inspected separately in each application.

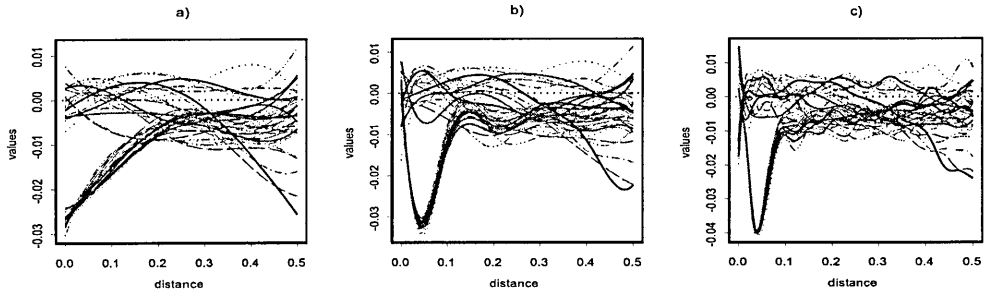

Figure 5.2: Smoothed L-functions with 5, 8, and 20 splines for 20 simulated random patterns and 20 simulated regular patterns

b) **Choice of clustering algorithm** A number of clustering algorithms were applied to the scores on the principal components when the method was used with groups of clustered and random patterns as well as random and regular patterns, including single linkage, complete linkage and Ward's method. Overall, Ward's method produced the most compact clusters and was thus chosen to be used in both the feasibility study and for the application. To illustrate this, Figure 5.3 shows the dendrogram resulting from a) the single linkage algorithm, b) Ward's method. Here, the FPCA method was applied to 20 simulated random patterns (ran1, ... , ran20 in the plot) and 20 simulated regular patterns (reg1, ... , reg20 in the plot) on the unit square. Since all patterns within each of the two groups were generated from the same algorithm, their spatial patterns should have very similar properties, resulting in very similar scores on the first principal components within the groups. This should be reflected in compact sub-trees in the dendrogram. It is clear that this is not the case for the dendrogram resulting from the single linkage algorithm whereas the dendrogram resulting from Ward's method yields two relatively compact groups. The complete linkage algorithm yields very similar results to Ward's method.

5.2.1 Separation performance and choice of summary statistic, (c) and (d)

As indicated above, the performance of the PCA was assessed in terms of its ability to distinguish between groups of patterns with distinct and known

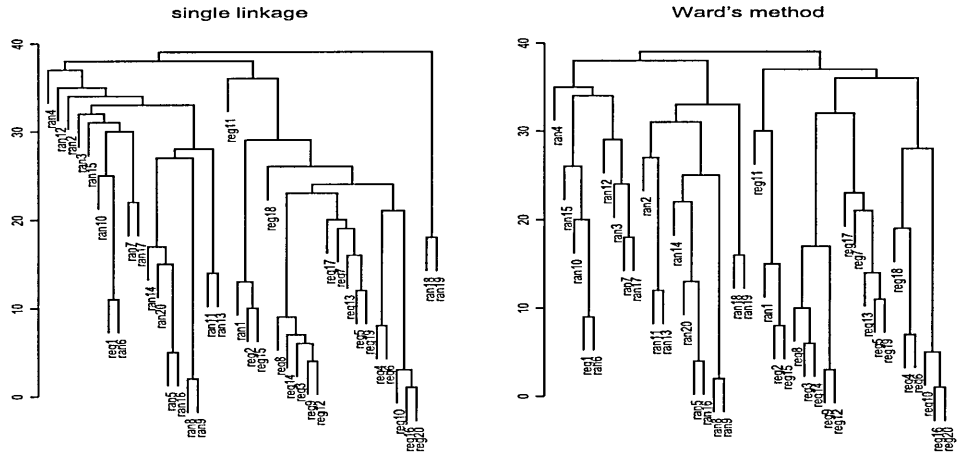

Figure 5.3: Dendrogram of scores on first two principal components using a) the single linkage algorithm and b) Ward's method

spatial behaviour. In this section we summarise the results of an extended simulation study where the performance was assessed separately for homogeneous and inhomogeneous patterns. Since the estimation of the second order summary statistics is more complicated in the inhomogeneous case, as the additional step of estimating the intensity is required, it is not obvious whether the performance is the same for both homogeneous and inhomogeneous patterns.

Unless otherwise stated all simulated patterns consisted of an expected number of $n = 100$ points in order to ensure that everything but the spatial patterning is kept constant. Nevertheless, the number of points in the different patterns should not have an influence on the performance of the method, since the second order summary statistics are standardised by the intensity of the pattern. Only if the number of points is small (i.e. < 20) problems

may occur since the estimated function gets very coarse.

5.2.1.1 Homogeneous case

We begin by assessing the ability of the functional principal component method to distinguish between a set of random patterns and a set of clustered patterns, as well as a set of random patterns and a set of regular patterns respectively. In a next step we assess the sensitivity of the method to distinguish patterns with different degrees of clustering and regularity, respectively. Finally, we consider clustered, regular and random patterns and the performance of the method in this context.

Clustered versus random patterns and regular versus random patterns In the simplest case, a set of 20 homogeneous Poisson cluster processes were simulated on the unit square with a parent process of intensity $c = 10$ and daughter processes with radius $rad = 0.025$. In addition, a set of 20 binomial processes, i.e. Poisson processes with a fixed number of points, were generated. L-functions as well as pair-correlation functions were estimated for each of the patterns. These were smoothed with 10 cubic B-splines and then subjected to a functional principal component analysis followed by a cluster analysis of their scores in the first two principal components. The method was capable of distinguishing the two groups of patterns perfectly: there was no misclassification, either with L-functions or with pair-correlation functions.

A similar result was achieved with two sets of 20 random and 20 regular (hard core) processes, i.e. patterns generated from a Strauss process with $\gamma = 0$; again there were no misclassifications.

Figure 5.4 shows the resulting dendrogram for both situations using the homogeneous L-function as a summary statistic. Figure 5.4 a) shows the dendrogram for the clustered versus random patterns and Figure 5.4 b) shows the dendrogram for the regular versus random patterns.

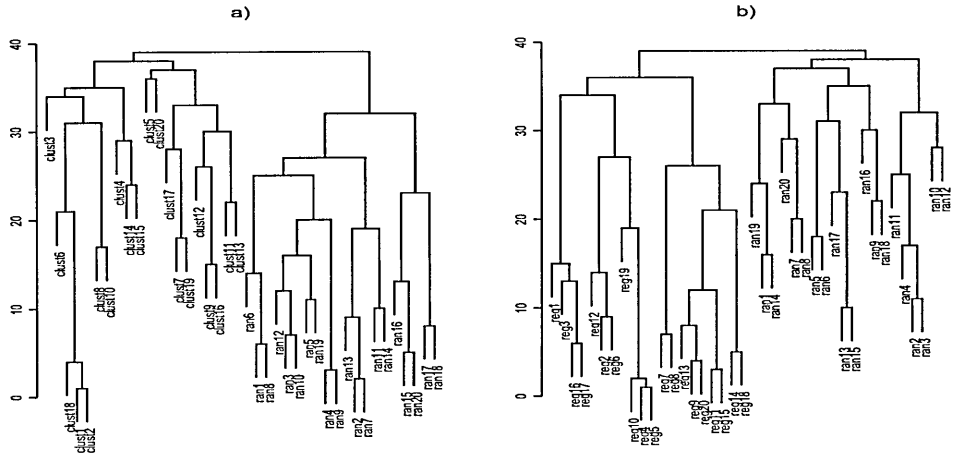

Figure 5.4: Clustered scores on the first two principal components for 20 regular and 20 random patterns using FPCA on L-functions (a) and clustered scores on the first two principal components for 20 clustered and 20 random patterns again using FPCA on L-functions (b).

Strength of regularity In order to assess the sensitivity of the method when the groups of patterns become increasingly similar, Strauss processes (Strauss 1975) with different levels of regularity were compared to hard core

function repulsion	weak ($\gamma = 0.9$)	medium ($\gamma = 0.5$)	strong ($\gamma = 0.1$)
L-	0%	0%	10 %
pair-correlation	0%	0%	5%

Table 5.1: Percentage of misclassified processes after FPCA on L-functions and pair-correlation functions when comparing a group of hard core processes (very strong repulsion) to a group of processes with different levels of repulsion.

processes. The interaction parameter was chosen as $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, where we have a hard core process if $\gamma = 0$ and a Poisson process if $\gamma = 1$. Each of the ten sets was compared to 20 simulated Strauss processes with complete inhibition (hard core process) within a radius of $rad = 0.05$. There was only a small number of misclassifications when repulsion was very strong, i.e. when the two groups of processes were very similar. See Table 5.1 for an overview of these results.

Strength of clustering For the purpose of assessing the sensitivity of the method with groups of different levels of clustering two sets of 20 Thomas processes with weak and strong clustering, respectively, were simulated. Here, the mean number of points per cluster was kept constant such that strength of clustering was interpreted as equivalent to size of clusters. Thus patterns with smaller clusters result in stronger clustering as the intensity of the points within the clusters is higher. Table 5.2 summarises the results for the different summary statistics. It is clear that the method again is very sensitive and manages to distinguish between the patterns very well, with the pair correlation function performing slightly better.

	percentage misclassified
L-function	5 %
pair-correlation function	0 %

Table 5.2: Percentage of misclassified processes after FPCA on L-functions and pair-correlation functions, respectively when comparing a group of weakly clustered processes to a group of strongly clustered processes

Clustered versus random versus regular patterns Three groups – 20 regular, 20 random and 20 clustered patterns – were simulated to investigate the performance of the method with regular, clustered and random patterns. We anticipated above (see Section 5.1.3) that it is likely that regular and random patterns will be classified as more similar than clustered and random patterns. We here compare the results for the suggested modification of the functional component analysis using the correlation matrix that we apply in this thesis to the original approach using the covariance matrix.

When using the L-function, 90.36% of the variation amongst the functions was explained by the first two PC's when the covariance matrix was used; this could be increased to 94.94% by using the correlation matrix. When the pair-correlation function was used, both the approach using the covariance matrix and the approach using the correlation matrix accounted for 99.5% of the variation. Figures 5.5 and 5.6 show that for both summary statistics, the three clusters were clearly identified by the method, and in both cases the regular patterns seemed more similar to the random ones, as predicted. The three clusters appear more distinct when the pair-correlation is used as a summary statistic. It seems advisable to use the FPCA on the correlation matrix whenever clustered, regular and random patterns are likely to

occur and the L-function is being used. Apparently, this difference is less pronounced when pair-correlation functions are used.

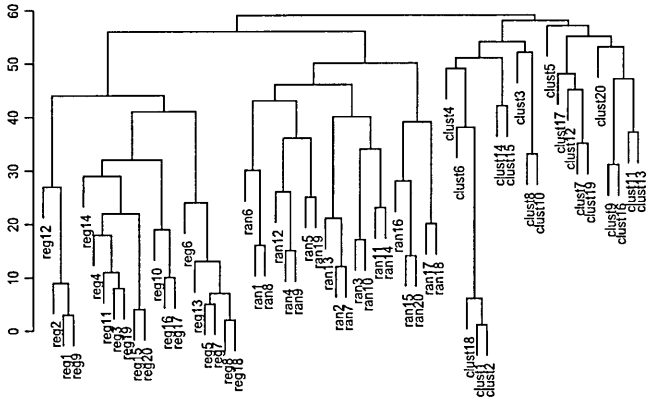

Figure 5.5: Clustered scores on the first two principal components for 20 clustered, 20 regular and 20 random patterns using FPCA on L-functions

5.2.1.2 Inhomogeneous patterns

In many applications, and in the case of the Cooljarloo data set, the assumption that the point pattern is homogeneous does not always hold and it is not clear whether the above results can be generalised to this case. Thus, in order to assess the performance of the method with inhomogeneous data the same comparisons as in Section 5.2.1.1 were repeated using simulated inhomogeneous patterns and the inhomogeneous L-function and pair-correlation function, respectively.

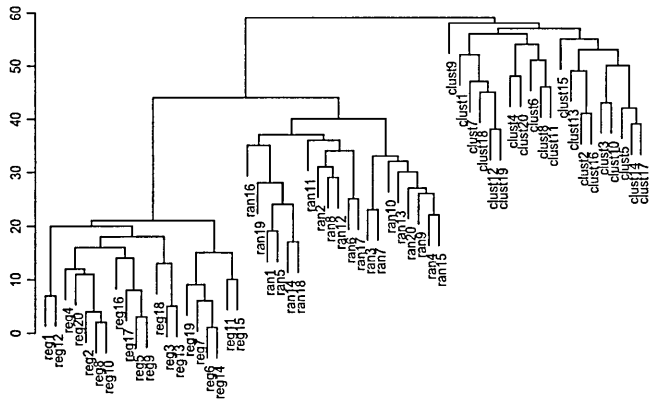

Figure 5.6: Clustered scores on the first two principal components for 20 clustered, 20 regular and 20 random patterns using FPCA on pair-correlation functions

Clustered versus random patterns and regular versus random patterns In a first step, 20 inhomogeneous Poisson cluster processes were simulated. These were Poisson cluster processes with an inhomogeneous Poisson process as a mother process. The intensity increased according to the trend function $f(x, y) = 30y$. These were compared to 20 inhomogeneous Poisson processes with the same trend function. When the inhomogeneous L-function (3.5) was applied, the method was strongly influenced by the bias in the estimation for larger distances (Baddeley et al. 2000). Restriction to smaller distances, e.g. by only considering distances up to 0.2, improved the results. These problems were clearly less pronounced when the pair-correlation function was used. In addition, there was no need to cut the pair-correlation functions to shorter distances. The results in Section 5.2.1.1 already indicated that the pair-correlation leads to a better perfor-

	% misclassified clustered, random	% misclassified regular, random
L-function, uncut	32.5 %	43.5 %
L-function, cut	2.5 %	6.5 %
pair-correlation function, uncut	0 %	0 %

Table 5.3: Percentage of misclassified processes after FPCA on L-functions and pair-correlation functions, respectively when comparing a group of inhomogeneous clustered processes to a group of inhomogeneous random processes and when comparing a group of inhomogeneous regular processes to a group of inhomogeneous random processes

mance of the FPCA method than the L-function. Apparently the difference in performance increases in the context of inhomogeneous patterns, with the pair correlation function showing optimal results and the uncut L-function yielding extremely unreliable results. See Table 5.3 for a comparison of the percentage of misclassifications[2] for L-functions that were not cut or cut to shorter distances, as well as for pair-correlation functions.

Similarly, 20 inhomogeneous Strauss processes were simulated. The intensity increased according to the trend function $f(x, y) = 30y$. These were compared to 20 inhomogeneous Poisson processes with the same trend function.

Again, there was a strong bias in the estimated L functions at larger distances, heavily influencing the performance of the method. Results improved when the pair-correlation function was used and again, there was no need to cut the pair-correlation functions to shorter distances. The approach yields even more unreliable results with a misclassification of 43.5 % than

[2]The number of misclassifications was determined as follows: 1) Identify the group with the largest number of patterns of one of the two types; 2) count the number of patterns of the other type allocated to this group, 3) add to this the number of patterns of the first type allocated to the other group. Note that this means that there cannot be more than 50% misclassifications.

repulsion	weak ($\gamma = 0.9$)	medium ($\gamma = 0.5$)	strong ($\gamma = 0.1$)
L-function, uncut	23%	33%	45 %
L-function, cut	5%	12%	25 %
pair-correlation f.	0%	0%	10%

Table 5.4: Percentage of misclassified processes after FPCA on L-functions and pair-correlation functions when comparing a group of hard core processes (very strong repulsion) to a group of processes with different levels of repulsion.

in the previous comparison when the L-function is used, whereas the pair-correlation function still performs flawlessly. See Table 5.3 for a comparison of the percentage of misclassifications for L-functions which were uncut and cut to shorter distances, as well as for pair-correlation functions.

Strength of regularity Again, the sensitivity of the method when the groups of patterns become increasingly similar was assessed using inhomogeneous Strauss processes with different levels of regularity and comparing these to hard core processes. The interaction parameter was chosen as $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, where we have a hard core process if $\gamma = 0$ and a Poisson process if $\gamma = 1$. Each of the ten sets was compared to 20 simulated inhomogeneous Strauss processes with complete inhibition (hard core process) with a radius $rad = 0.05$. For the L-function results are again poor, unlike for the pair correlation function, where there again was only a small number of misclassifications when repulsion was very strong, i.e. when the two groups of processes were very similar. See Table 5.4 for an overview of these results.

	percentage misclassified
L-function, uncut	45 %
L-function, cut	30 %
pair-correlation function, uncut	0 %

Table 5.5: Percentage of misclassified processes after FPCA on L-functions and pair-correlation functions, respectively, when comparing a group of weakly clustered processes to a group of strongly clustered processes

Strength of clustering When two sets of Poisson cluster processes with different levels of clustering (as in Section 5.2.1.1) were compared, the differences between the two summary statistics were even more pronounced, see Table 5.5. The pair-correlation function yields a result with no misclassifications whereas even the cut L-function leads to misclassifications of 30%.

Clustered versus random versus regular patterns Again, three groups of inhomogeneous patterns – 20 regular, 20 random and 20 clustered processes – were simulated to investigate the performance of the method with regular, clustered and random patterns.

When using the L-function (cut to close distances), 63.4% of the variation amongst the functions could be explained by the first two PCs when the covariance matrix was used, which increased to 72.49% by using the correlation matrix. When the pair-correlation function was used, the approach using the covariance matrix accounted for 83.5% of the variation and the approach using the correlation matrix accounted for 87.3% of the variation.

Again, as in the homogeneous case the three clusters were clearly identified by the method when the pair correlation function was used. There were a few misclassifications, though, when the L-function was used (cut to close

distances). In both cases the regular patterns seemed more similar to the random patterns than the clustered patterns, as predicted.

Conclusion These results clearly indicate that it is advisable to use the pair-correlation function for the FPCA whenever an inhomogeneous second-order summary statistic is used. We will take this aspect into account when applying the functional principal component method to the Cooljarloo data set in Section 5.4 since most of the subpatterns in the data set have been found to be inhomogeneous (see Section 4.1.2.3).

5.3 Simulation study: erroneous data

The performance of the approach when applied to noisy data is assessed, since data used in ecological studies are susceptible to error, in particular when data collection is challenging, as for example in Armstrong (1991) where the location of more than 6000 individuals had to be determined. Due to this large number of plants and also the fact that it is difficult to determine the exact location of plants with, e.g. a creeping habit, errors are very likely. Identifying the exact location of a plant and distinguishing individuals may be a complex task in itself (Magurran 1988; Stoll and Weiner 2000). Furthermore, the data in Armstrong (1991) were collected on a grid with cells marked to 1m $\times$ 1m but estimated and recorded on a 10cm $\times$ 10cm grid and the location was only accurate up to the fineness of the grid.

In order to investigate the robustness of the method, three different types of error common to ecological applications are considered:

a) The location is inaccurately recorded due to human error or technical problems.

b) The location is recorded on a grid. This grid is fine enough that the probability that any resulting cell contains more than one data point is very small. Nevertheless, strictly speaking the recorded location does not reflect the exact location of the individual plants.

c) The wrong species is recorded, i.e. marks were accidentally confounded.

The simulation study mimicked these three cases by generating point patterns of different spatial behaviour and subsequently increasing the noise in the data. For a) an increasing degree of random noise is deliberately added to the original locations; for b) the data are discretised to different resolutions; and for c) an increasing number of marks from one species are randomly replaced by marks of another species. In all three cases, the focus is to identify the degree of distortion that may be introduced whilst still preserving sufficiently accurate results, in order to provide ecologists with an indication as to the degree of accuracy in the data collection required[3].

As above, all simulated patterns consisted of an expected number of $n = 100$ points, unless otherwise stated, in order to ensure that everything but the spatial patterning is kept constant. Nevertheless, the number of points

[3]Note that non-registration of an individual may be another common error. We will here assume that this mistake happens independent of the location of the individual and can thus be regarded as independent thinning. As indicated in Section 3.3.2, the summary statistics are invariant under independent thinning such that we do not consider this issue here any further. This assumption may be disputable as the probability of not registering an individual may vary across the plot with changing light conditions, say. However, in the study data set no additional information is available which could be used to identify areas where non-registration is more likely.

in the different patterns should not have an influence on the performance of the method, since the second order summary statistics are standardised by the intensity of the pattern.

Also, as above, each set of patterns consisted of 20 patterns. I.e. when two sets of patterns are considered the method is applied to a total of 40 patterns.

5.3.1 Inaccurate location

In order to mimic erroneously recorded locations, sets of clustered, random and regular point patterns were generated as described in Section 5.2. Subsequently, values from a normal distribution $N(0, \sigma^2)$ were generated and added to the original x- and y-coordinates. Here, the strength of noise is reflected in the size of σ^2.

Note that after this procedure the patterns generated from a homogeneous Poisson process will still show complete spatial randomness whereas the clustered and regular processes will become increasingly similar to the Poisson processes.

Figures 5.7 (a) and (c) show the percentage of misclassifications for different degrees of noise for inhomogeneous patterns, for simulations where clustered versus random and regular versus random patterns were generated as in Section 5.2 and analysed using the pair-correlation function. Figure 5.7 (b) and (d) show the analogous results for the L-function.

Similarly, Figures 5.8 (a) and (c) show the percentage of misclassifications for different degrees of noise for inhomogeneous patterns, for simulations where clustered versus random and regular versus random patterns were

generated and analysed using the pair-correlation function. Figure 5.7 (b) and (d) show the analogous results for the L-function.

In the homogeneous case, for the analyses with the L-function, results become increasingly unreliable from a standard deviation of 0.2 onwards for clustered patterns and from 0.03 for regular patterns. Since the patterns have been generated on the unit-square, 0.2 is equivalent to 20% of the plot length, and 0.03 to 3% of the plot length. In the inhomogeneous case, for the analyses with the L-function, results become increasingly unreliable from a standard deviation of 0.15 onwards for clustered patterns and from 0.025 for regular patterns. Since the patterns have been generated on the unit-square, 0.15 is equivalent to 15% of the plot length, and 0.025 to 2.5% of the plot length.

Translated into practical terms, this means that for a data set as in Armstrong (1991) where the locations have been recorded on a
22 m $\times$22 m plot, the 2σ region for the noise has a width of 1.1 m for regular patterns and 6.60 m for clustered patterns. The results are similar but slightly better for the pair-correlation, with results getting increasingly unreliable from a standard deviation of 0.22 onwards for clustered patterns and from 0.03 for regular patterns (i.e. 9.68m and 1.32m in the study data plot).

5.3.2 Data collected on a grid

Spatial point process models assume that the location of objects under investigation was recorded on a continuous scale. In practice, however, this is

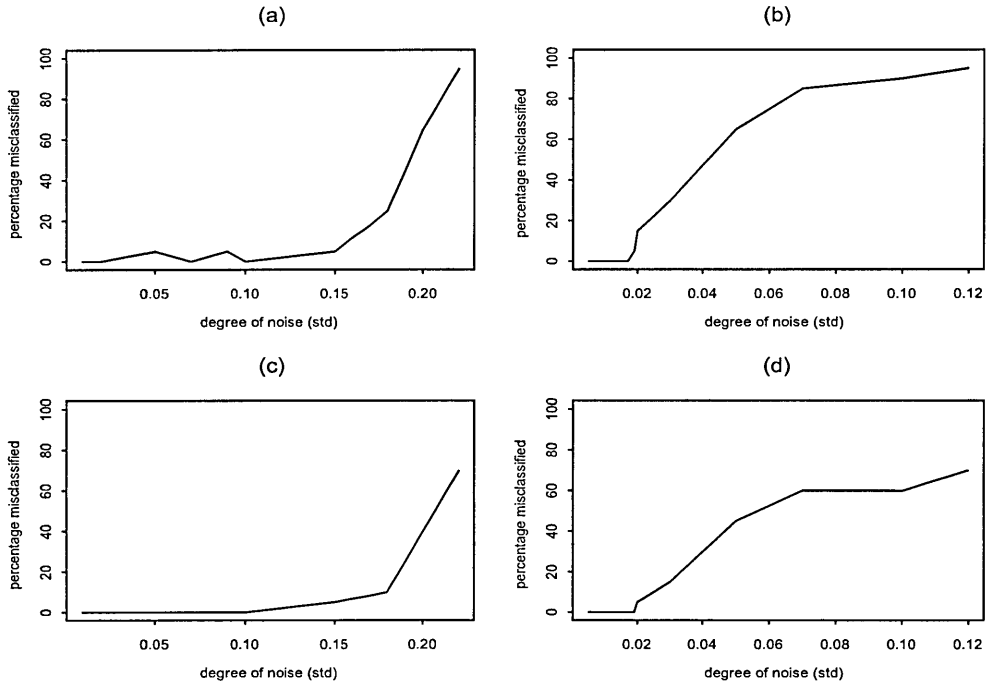

Figure 5.7: Percentage of misclassifications as a function of the strength of noise, homogeneous patterns. (a): clustered versus random patterns, pair-correlation function, (b): regular versus random patterns, pair-correlation function, (c): clustered versus random patterns, L-function, (d): regular versus random patterns, L-function

not the case since data are typically sampled from a discretised space due to real world constraints. Ecologists typically record a plant's location on a grid, e.g. Armstrong (1991). This grid has a very fine resolution such that it is very unlikely for any two points to appear in the same grid cell but this is not impossible. Here, we generate patterns as in Section 5.2 and then modify the location to an increasingly coarse grid by rounding the coordinates.

Figure 5.9 shows the number of misclassifications resulting from an increasingly coarse grid when the simulated patterns are homogeneous. Figure

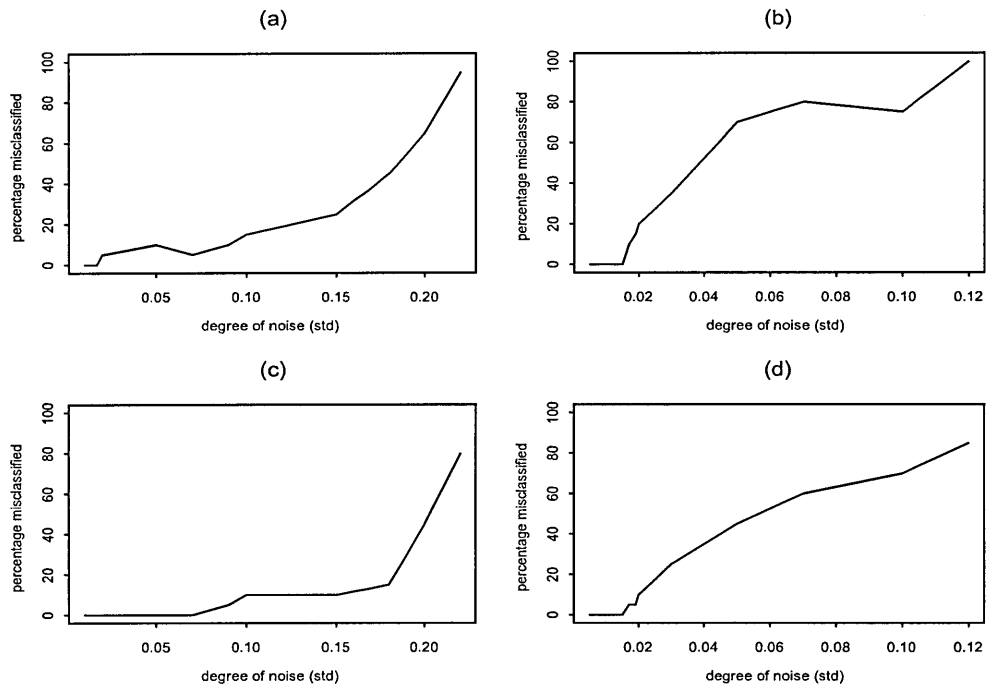

Figure 5.8: Percentage of misclassifications as a function of the strength of noise, inhomogeneous patterns. (a): clustered versus random patterns, pair-correlation function, (b): regular versus random patterns, pair-correlation function, (c): clustered versus random patterns, L-function, (d): regular versus random patterns, L-function

5.9 a) and c) show the results when the L-function was used for clustered and regular patterns, respectively. Figure 5.9 b) and d) show the analogous results when the pair-correlation function was used.

The results are very similar, for both regular and clustered patterns as well as for the two summary statistics. Misclassifications only occur when the grid becomes as coarse as consisting of $4 \times 4 = 16$ cells. Hence, the method only fails when the grid has been extremely coarse-grained, so coarse that this choice of grid would be highly impractical in applications. In order to

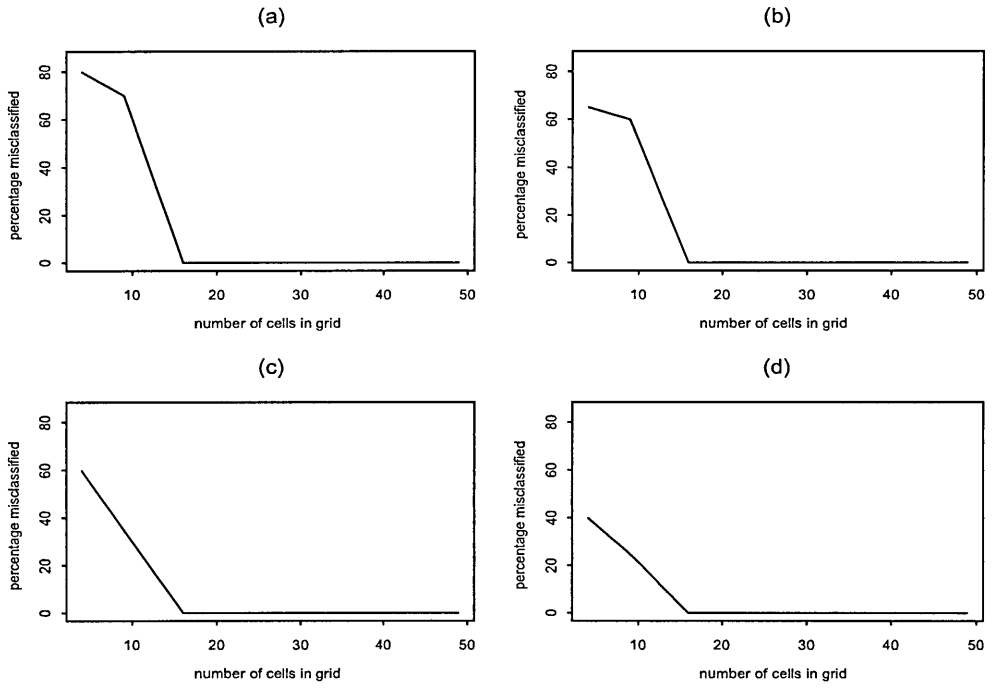

Figure 5.9: Percentage of misclassifications as a function of the number of gridcells used (homogeneous patterns). a): clustered patterns, L-function, b): clustered patterns, pair-correlation function, c): regular patterns, L-function, d): regular patterns, pair-correlation function

measure the exact location of 20 sets of approximately 100 points each, a grid of this coarseness would not be chosen. Also, note that, theoretically, spatial point processes are defined as simple processes where the probability of two points occurring at the same location is zero (Møller and Waagepetersen 2003b). Using a very coarse grid will yield a large number of points recorded as occurring in the same location which would make the use of spatial point pattern modelling invalid.

However, the finer structure within the clusters is lost leading to less detailed

results and changing the relative differences between species. Thus, the relative position of the individual patterns in the hierarchy is both obscured and altered. To illustrate this, Figure 5.10 shows the resulting dendrograms (analysis done using L-functions and homogeneous clustered versus random patterns) for a) the original data, b) data on a 15×15 grid and c) data on a 5×5 grid. Despite the increasing coarseness none of the subpatterns is misclassified but the structure within the clusters changes. For example "ran18" and "ran19" appear very similar in Figures 5.10 a) and b) but have moved rather far away from each other in Figure 5.10 c). This change is stronger within the group of random patterns. Similar effects can be observed when the pair-correlation function is used and also for inhomogeneous patterns.

Similarly, Figure 5.11 shows the number of misclassifications resulting from an increasingly coarse grid when the simulated patterns are inhomogeneous. Figure 5.11 a) and c) show the results when the L-function was used for clustered and regular patterns, respectively. Figure 5.11 b) and d) show the analogous results when the pair-correlation function was used. Here, subpatterns are misclassified when the grid consists of $6 \times 6 = 36$ cells, which still is a degree of coarseness which would not be used in practice.

On the whole, we conclude that the fact that the data have been recorded on a grid does not meaningfully influence the overall results but some changes in the finer structure of the results can be observed if the grid becomes increasingly coarse.

5.3.3 Wrong species recorded

In practice, human error might lead to the data collector recording the wrong species. This can be the result of a number of mistakes, such as misrecorded species and other data processing errors. Most of these problems are not systematic and thus cannot be meaningfully simulated and so will not be considered here. We consider a situation where individual plants from one specific species have been erroneously recorded as being another specific species with a different spatial behaviour. We assume it is unlikely that it is an arbitrary misclassification, but one with a species of similar phenotypic characteristics as we assume that this is the most likely error that may happen.

In order to mimic this situation, sets of regular and random as well as clustered and random patterns were simulated as described above (see Section 5.2). Subsequently, individuals from one of the clustered or regular patterns, respectively, were randomly labelled with the label of one of the random patterns. The probability of this relabelling was varied and the effect of an increasing probability of misclassification investigated.

To assess the influence of misclassification we consider the following. For two point patterns x and y let $d(x, y)$ be the Euclidean distance between the score vectors of the corresponding first two principal components. For two sets of point patterns $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_m\}$ we can now calculate the average of these distances for pattern x_i to the other patterns in its group relative to the average distance of pattern x_i to all other patterns, i.e. we use the relative distance rd as defined by

$$rd(x_i, x, y) = \frac{(n-1)^{-1} \sum_{j=1}^{n} d(x_i, x_j)}{(n-1)^{-1} \sum_{j=1}^{n} d(x_i, x_j) + m^{-1} \sum_{k=1}^{m} d(x_i, y_k)}$$

and assess the influence of the method on rd, assuming that a pattern will be classified into the nearest group. If the distance to the original group is larger than the distance to the wrong group, i.e. when $rd(x_i, \mathbf{x}, \mathbf{y}) \geq 0.5$, the modified pattern is more likely to be wrongly classified.

To illustrate this, Figure 5.12 shows a plot of how an individual subpattern slowly "moves" from its own group of regular patterns into the other group of random patterns with increasing probability of an individual point in a subpattern being re-labelled (see circle).

Figure 5.13 shows the results from 100 simulations for regular patterns (Strauss processes with complete inhibition) versus random patterns, using the L-function. $rd(x_i, \mathbf{x}, \mathbf{y})$ is plotted as a function of the probability of an individual point in a subpattern being re-labelled.

From a probability of re-labelling of more than 0.24 onwards $rd(x_i, \mathbf{x}, \mathbf{y}) \geq 0.5$ in the case of regular versus random patterns when L-functions are being used. It increases to 0.27 for the pair-correlation function in the same setting, and to 0.32 and 0.36 in the of case clustered versus random patterns, for L-function and pair-correlation function, respectively.

Note that the problem of misclassification has not been assessed for inhomogeneous patterns. Whereas the simulation and estimation process for the 100 simulations could be automated in the homogeneous context, this could not be done in the inhomogeneous context since, as noted previously, the estimation of both the inhomogeneous L-function and the inhomogeneous pair-correlation function requires an estimation of the intensity surface which

in turn involves the choice of a bandwidth. This choice cannot be done automatically as the appropriate degree of smoothness depends on the specific application, making the entire simulation highly impractical.

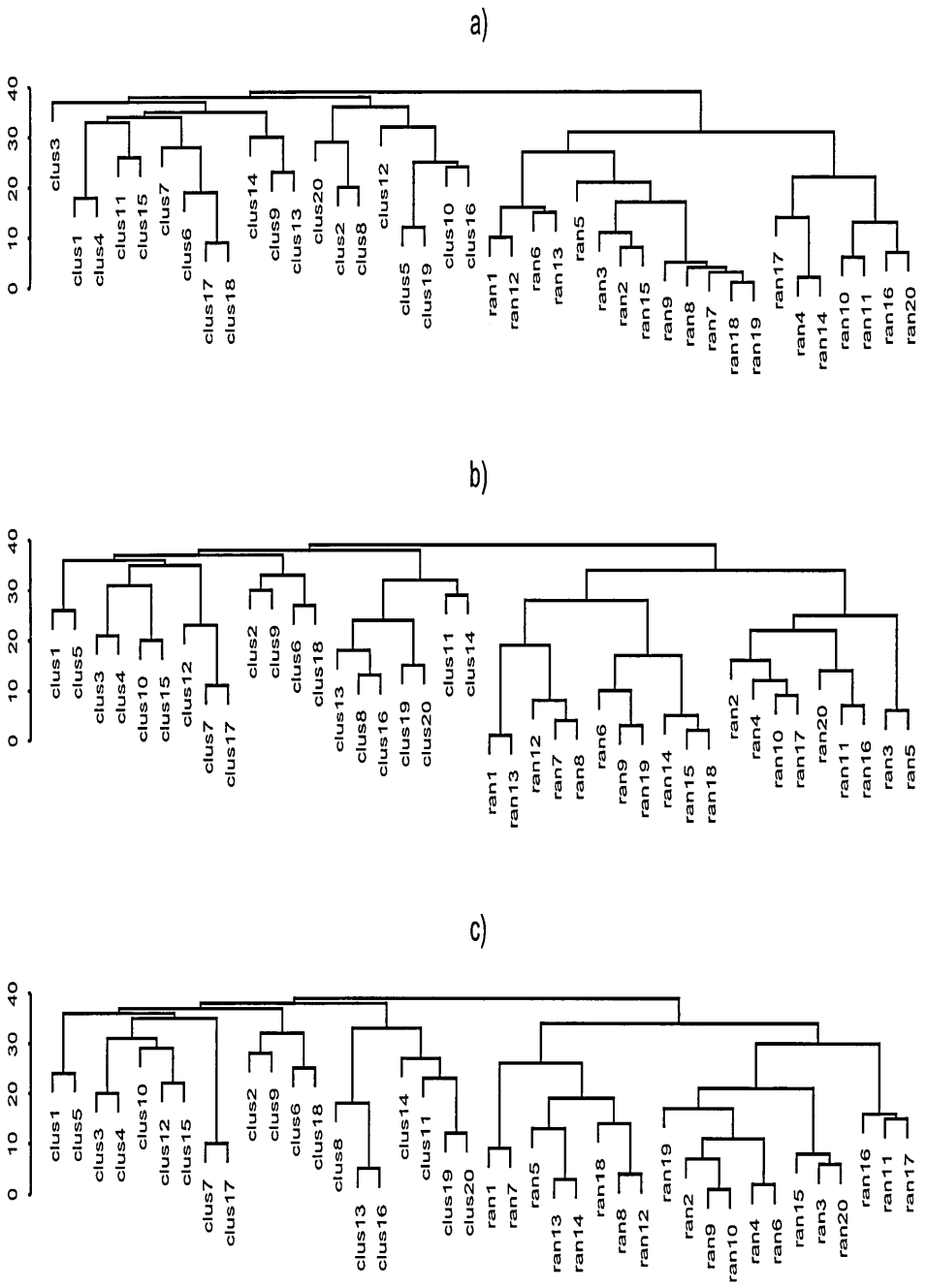

Figure 5.10: Dendrogram of scores on the first two principal components from FPCA on L-functions for homogeneous clustered versus random patterns, for original data a), data on a 15×15 grid b) and data on a 5×5 grid c)

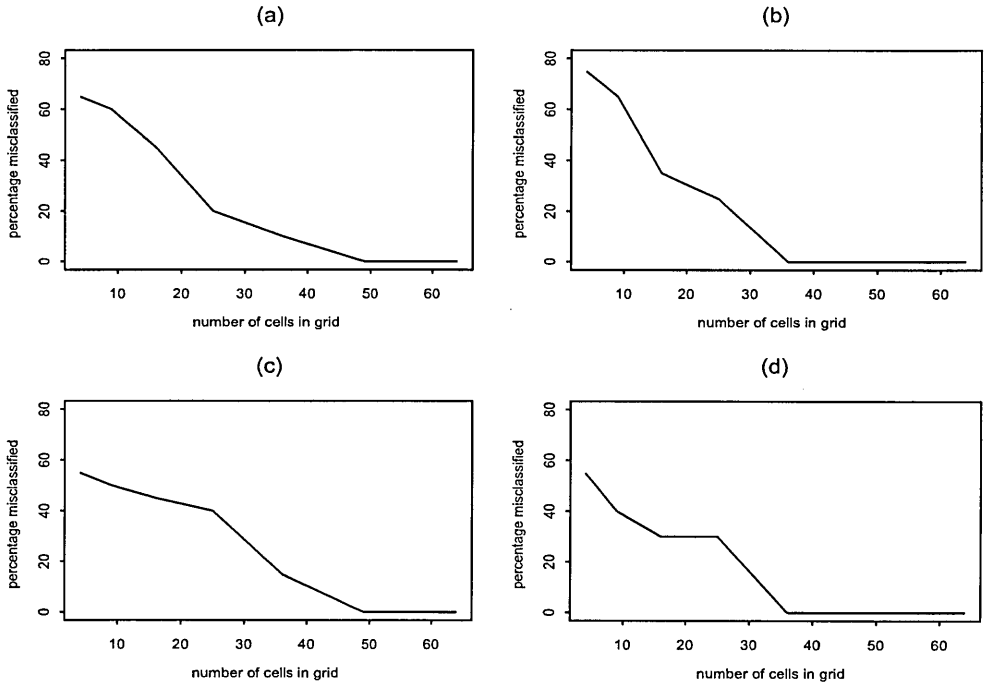

Figure 5.11: Percentage of misclassifications as a function of the number of gridcells used (inhomogeneous patterns). a): clustered patterns, L-function, b): clustered patterns, pair-correlation function, c): regular patterns, L-function, d): regular patterns, pair-correlation function

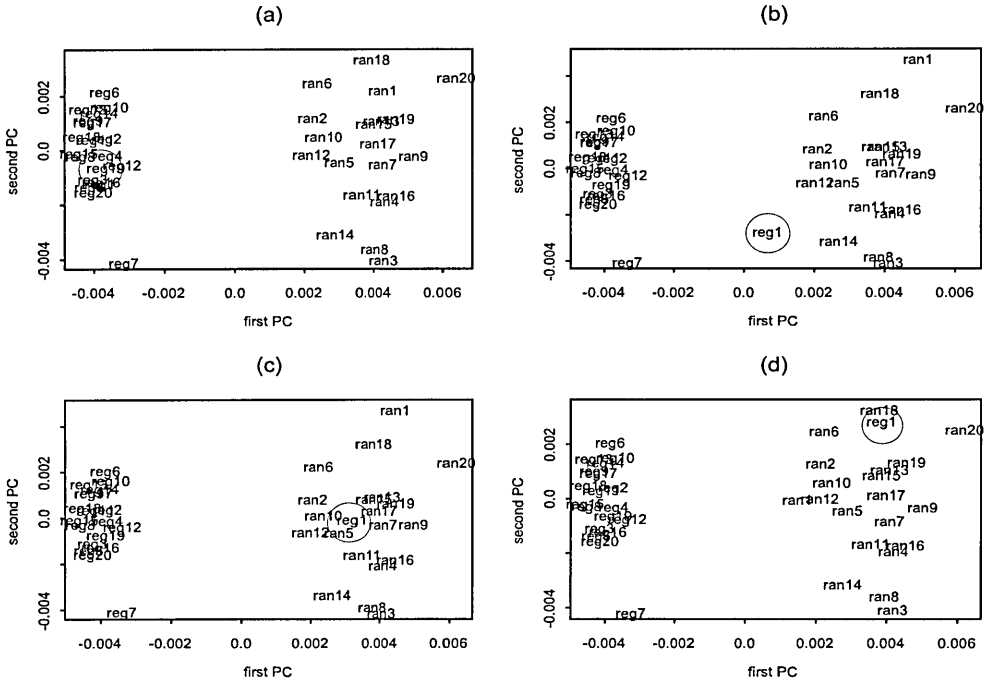

Figure 5.12: Demonstration of the regular subpattern "reg1" (see circle) becoming increasingly dissimilar from the group of regular patterns with increasing probability of an individual point in a subpattern being re-labelled; relabelling probabilities are a) 0.0, b) 0.3, c) 0.7 and d) 0.9.

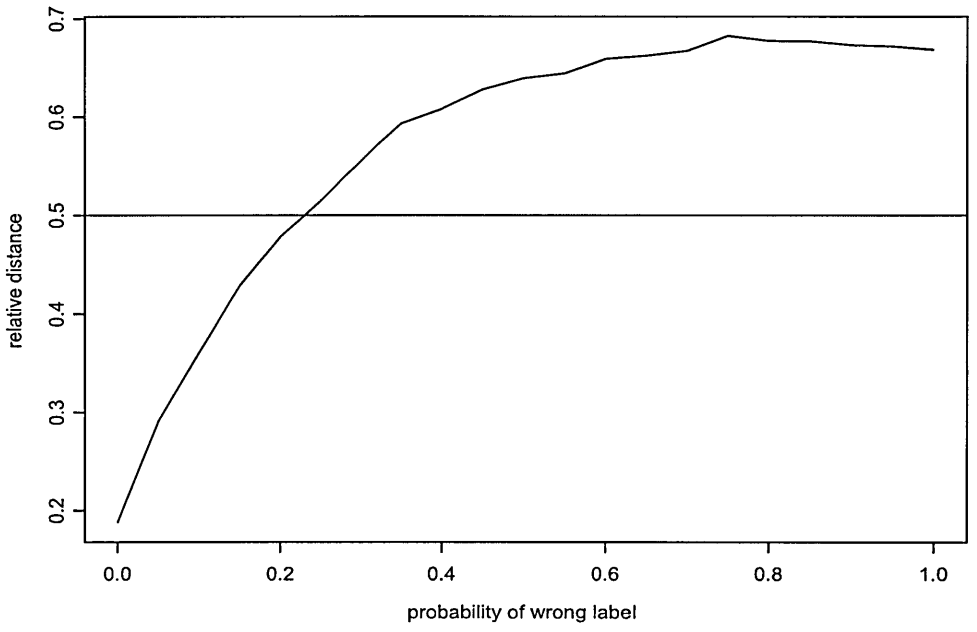

Figure 5.13: $rd(x_i, \mathbf{x}, \mathbf{y})$ as a function of the probability of an individual point in a subpattern being re-labelled for regular (Strauss process with complete inhibition) and random patterns using the L-function

5.4 Application to data set

Informed by both the feasibility study and the simulation study for erroneous data we can now apply the method to the Cooljarloo data set. Both studies have consistently shown that the performance of the pair-correlation function is much better than that of the L-function, in particular when the data are inhomogeneous. Since most of the individual species show strong inhomogeneity the inhomogeneous pair correlation function (Baddeley et al. 2000) was used to analyse the Cooljarloo data set. Furthermore, as indicated in Section 5.3.1 the method would become unreliable if the location had been incorrectly recorded by 9.68 m for clustered patterns and by 1.32 m for regular patterns. Naturally, it is impossible to verify the exactness of the data collection several years later but it is highly unlikely that mistakes of this magnitude have been made, given that the data were collected on a grid with cells marked to 1m $\times$ 1m and estimated to an accuracy of 10 cm by 10 cm cells. In addition, this is a very fine grid, and so, according to Section 5.3.2, the results should not be very different from those that would have been yielded with data collected on a continuous scale. Naturally, we cannot assess the probability of a misclassification of the recorded individuals, neither from within the plot or from outside.

The estimated pair correlation functions were smoothed using 10 cubic B-splines. See Figure 5.14 for a plot of the smoothed pair correlation functions.

The first two principal components explain 84.9 % of the variance, the third principal component would only explain another 7.8 %. As before, we will thus restrict the further analysis to the first two PCs. The analysis of

the first two principal components (Figure 5.15) reveals that the first PC describes clustering at close distances and the second PC clustering at larger distances. Hence the most distinctive feature in the spatial behaviour is the species' behaviour at close distances.

Figure 5.16 shows the result of the hierarchical cluster analysis (Ward's method) on the scores of the first two principal components. Four different groups of patterns can be identified as indicated in the plot. The numbers represent the species numbers as given in Table 2.1.

Figure 5.17 plots the scores on the first two principal components. The grouping derived from the dendrogram is indicated by the same colours that were used in Figure 5.16. Blue numbers refer to group 1, red numbers refer to group 2, black numbers to group 3 and green numbers to group 4.

To interpret this grouping, we need to refer back to the principal components in Figure 5.15. For example, the second group (red numbers) scores low (negative) on the first principal component which represents clustering at close distances. All species in this group have a negative score in the first PC, indicating no clustering at close distances but some degree of repulsion, maybe a hard core. The second group (red numbers) also scores low on the second principal component, which represents clustering at larger distances. This indicates that the patterns in this group show no clustering at larger distances, ranging from random behaviour at larger distances (species 61 for example) to some repulsion at larger distances (species 20 for example)[4].

[4]Note that the repulsion observed for these species in this analysis do not contradict the results in Chapter 4. The methods applied there tested for significant deviations from CSR and found no case of a significantly repulsive pattern.

We will now look at the grouping within the individual groups in more detail and provide explanations for the spatial behaviour of the species in the groups.

- **Group 1 (blue):**

  Group 1 consists of species 3, 5, 6, 18, 25, 36, 48, 50, 51. All species in group 1 have negative scores on the first principal component indicating some repulsion at close distances (up to 20cm). Species 3, 18, 50 and 51 have a score of almost zero on the second principal component, indicating randomness beyond 20cm from each plant. This can be explained by the fact that the root system of the parent plant inhibits survival of daughter plants at close distances but not beyond (Armstrong 1991). Species 6 exhibits some light repulsion at further distances but on the whole behaves very similar to 50 and 51. Species 5, 25, 36 and 48 show increasing levels of clustering at distances beyond 20cm.

- **Group 2 (red):**

  The species in group 2, i.e. species 20, 33, 37, 42, 45, 47, 49 and 61 score negatively on the first principal component. Since the scores have larger negative values than those in group one, repulsion is stronger in this group. This means they all exhibit strong repulsion at close distances (up to 25cm). Some of the species within this group (47, 49 and 61) have a score around zero on the second principal component indicating randomness at larger distances, the other species show an increasingly negative score on the second, indicating different levels of repulsion up to a distance of approximately 30cm. All species in this group are

long-lived (Armstrong 2005). The individual plants have apparently established an exclusion zone around themselves over the years.

- **Group 3 (black)**:

  This is the largest and most heterogeneous group consisting of species 8, 12, 13, 14, 15, 26, 32, 34, 38 and 64. These species show increasingly positive scores on the first principal component indicating increasing levels of clustering at close distances (up to 20cm), starting from almost random for 12, 32, 34, via some light clustering for 8, 15 and 38 to stronger clustering for 13, 14, 26 and 64. On the second principal component we find negative, near zero and positive scores, with species 12, 13, 14, 26, 57 showing repulsion at distances beyond 20 cm, species 15, 32, 34 and 57 showing randomness and species 8 and 64 exhibiting clustering also between 20 and 30 cm distance.

- **Group 4 (green)**:

  The three species 19, 23 and 54 all show strong clustering at close distances. Both 19 and 54 show repulsion at distances beyond 20cm whereas species 23 shows strong clustering also beyond 20cm. All the species in this group are seed producing (Armstrong 2005); the clustering may indicate that they shed their seed mainly in close distances around them.

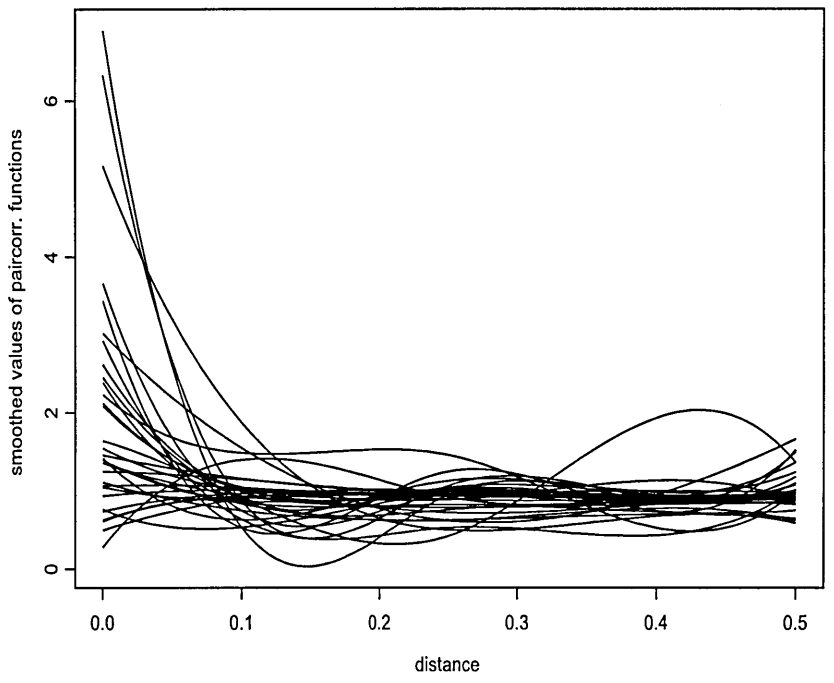

Figure 5.14: Plot of smoothed pair correlation functions for Cooljarloo data set

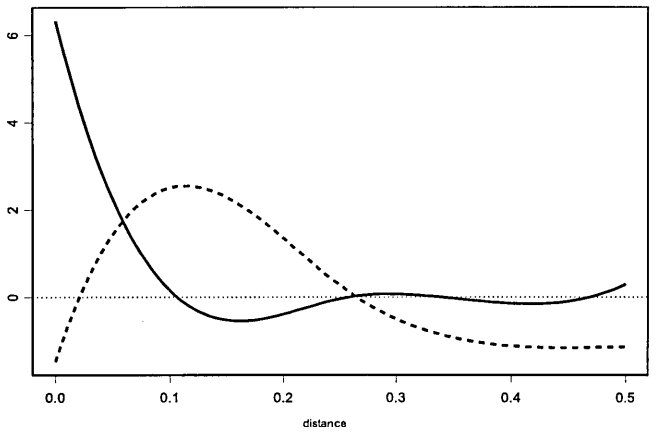

Figure 5.15: Plot of the first two principal component functions for Cooljarloo data set; the full line is the 1st PC, the dotted line is the 2nd PC

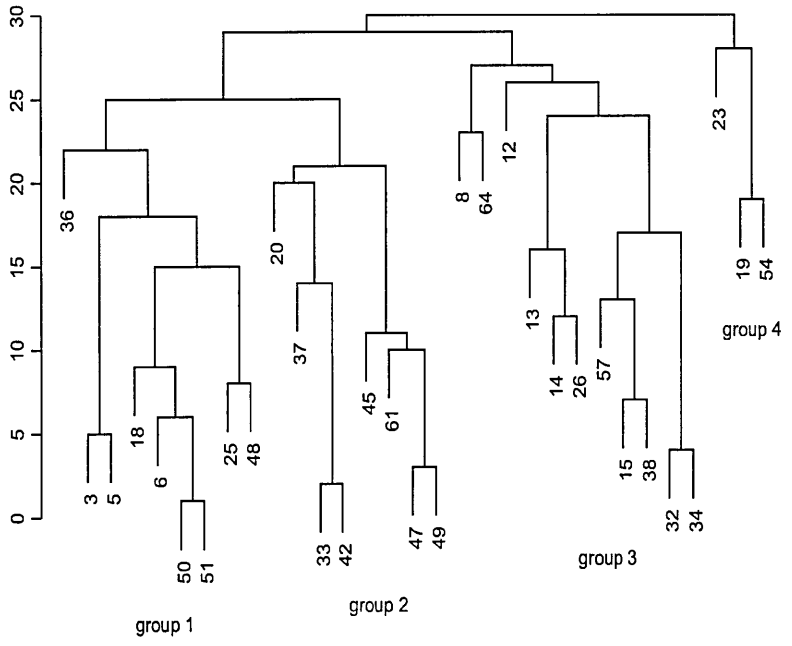

Figure 5.16: Dendrogram for the Cooljarloo data set after cluster analysis (Ward's method) of the scores on the first two principal components; numbers indicate species numbers as in Table 2.1

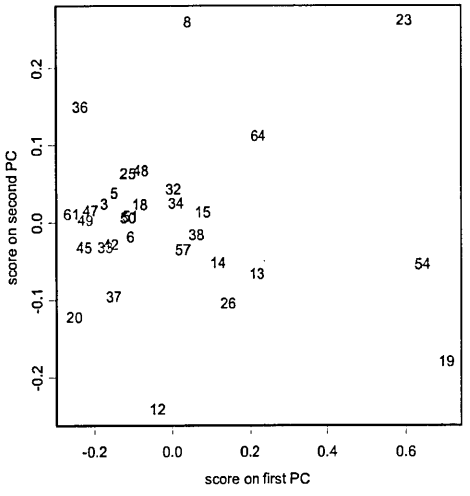

Figure 5.17: Plot of the scores on the first two principal components for Cooljarloo data set; numbers indicate species numbers as in Table 2.1

5.5 Discussion

5.5.1 Discussion of methodology

We present an approach to allow principal component analysis of spatial point patterns by applying functional principal component analysis to the second-order summary statistics of multi-type spatial point patterns. This yields a classification of the subpatterns into groups of similar spatial arrangement.

A feasibility study revealed that the method was capable of distinguishing clearly between both clustered versus random, and regular versus random patterns. Furthermore, the method proved to be sensitive enough to distinguish between similar types of patterns, i.e. patterns with a very similar degree of regularity. Overall, the results were very similar for L-functions and pair-correlation functions, with slightly better results for the pair-correlation function in some situations. This might have been a result of the cumulative nature of the L-function.

In a setting where regular, random and clustered patterns are present, the regular patterns appear more similar to the random patterns than do the clustered patterns, and as a consequence are more difficult to distinguish from the random patterns. Here again, the pair-correlation function produced slightly better results with a clearer classification. However, the observed similarity between regular and random arises as an inherent property of the patterns themselves. Whilst there is a finite limit to the spacings and inhibition associated with regular patterns, there is no limit (in principle) to the spacings and attractions associated with clustered patterns. Therefore there may be a "tending to unbounded" difference between random and

clustered patterns, whereas the regular patterns are constrained to complete inhibition. Consequently, when interpreting the results from a principal component analysis this must be kept in mind when comparing the strength of clustering with respect to the strength of regularity.

A detailed simulation study investigated the performance of the method in the presence of noise typical of ecological data. When random noise was added to the location of individuals, the method proved to be very stable. Only strong degrees of noise led to serious misclassifications. Similarly, when data were discretised into a grid of increasing coarseness, only a very coarse grid prohibited the overall classification of the subpatterns into the largest groups. However, the finer similarity structure became lost as a result of less precise data. Finally, the probability of an individual species being classified into a group with different spatial behaviour was only high when the probability of the species to be misidentified was at least 0.24.

In all three cases, again, the pair-correlation function yields slightly better results. Overall, the results from the study enable us to inform the applied researcher about the degree of noise which will cast doubt on the analysis. Using the results from the extended simulation study described here we are able to assess the reliability of the results for the data set. This could assure us that given the quality of the data collection described in Armstrong (1991), an influence of errors in the data collection other than profound species misclassification is very unlikely

Due to the ongoing technological development, larger numbers of similar data sets, as described in Burslem et al. (2001), will become available. Hence, there is much room for an extension of the approach, for example by

incorporating interspecies interactions based on two-dimensional L-functions or pair-correlation functions, or by incorporating marks into the analysis using a mixed (i.e. a combination of functional and non-functional) principal component analysis.

5.5.2 Discussion of biodiversity theory in this context

The functional principal component analysis revealed groups of species with similar spatial behaviour and established those aspects of spatial behaviour that vary most among the species in this data set. It might be that the spatial behaviour at small distances constitutes a niche-behaviour, i.e. a behaviour that the species have specialised in to facilitate coexistence. To understand this phenomenon better it is necessary to apply the method to other similiar data sets e.g. the one described in Enright et al. (2004), to establish whether similar results can be found for these. Furthermore, a similar analysis could be applied to the rainforest data sets collected through the CTFS network (see Section 2.3.2) to investigate the situation in a very different ecosystem.

In addition, the method may aid validating the "Janzen-Connell" hypothesis (see Section 2.3.2). The hypothesis predicts that plant populations become more segregated as they age, i.e. that older plants have a more regular pattern than younger plants. A FPCA of the pattern of plants of different age classes (as found on the CTFS plots) could reveal, whether plants of the same species but of different age are classified into different groups with regard to the spatial patterning.

5.5.3 Discussion of results for data set

When applied to the data set the method yielded a description of the most distinctive features in the spatial behaviour within the community, i.e. presence or absence of clustering at close distances (Illian et al. 2004, Illian et al. 2005). The second most distinctive feature was the presence or absence of clustering at larger distances. This may indicate that multi-species coexistence is facilitated through a strong variation in strength of intra-specific attraction in an environment as poor in nutrients as the heathlands of Western Australia (Beard 1990; Dixon 2005). However, it will be necessary to apply similar methods to other data sets in different and similar environments and with different degrees of biodiversity to verify this hypothesis.

The grouping of the species into groups of similar spatial behaviour does not coincide with other classifications such as regeneration method, genus or growth habit. To illustrate this, Table 5.6 shows the species sub-divided by the grouping resulting from the FPCA and also by regeneration method[5]. Apart from the very small group IV all groups contain both resprouters and seeders. Fisher's exact test was performed on these results in order to test whether there is an association between the FPCA grouping and the regeneration strategy. It yielded a non-significant result ($p = 0.3106$), i.e. there is not enough evidence that grouping and regeneration strategy are associated. Similarly, being of the same genus does not coincide with the grouping. There are cases where species from the same genus have been

[5]Note that the grouping in this Table 5.6 is different from the grouping obtained in Section 4.3. This is due to the fact that the grouping here is based on the species' position relative to the first two principal components rather than with regard to the categories 'clustered", "random" and "regular" or "homogeneous" and "inhomogeneous".

classified into the same group, such as *Hibbertia crassifolia* and *Hibbertia subvaginata* (both in group III) or *Banksia attenuata* and *Banksia menziesii* (also both in group III). However, there are also cases where species from the same genus fall into different groups, such as *Leucopogon striatus* (group II) and *Leucopogon constephioides* (group I) or *Hibbertia* species (group III) and *Hibbertia hypericoides* (group I).

As far as the growth habit is concerned, again different types of growth habit do not coincide with the grouping resulting from the FPCA. The abbreviation in Table 5.6 indicate the different types of growth habits found in the plants in this data set (Armstrong 2005). Again, Fisher's exact test was performed to test whether there is an association between the FPCA grouping and the growth habit. It yielded a non-significant result ($p = 0.4428$), i.e. grouping and growth habit are not associated.

This phenomenon might be due to the fact that we are analysing a very ancient community where over millions of years the spatial behaviour has been adapted across families to enable coexistence (Dixon 2005). This may indicate that a niche specification with regard to spatial clustering at close distances may have taken place, thus providing evidence in favour of the niche specification hypothesis (see Section 2.3.1). In order to verify this, further studies would have to be undertaken to compare results across different communities with different degrees of biodiversity.

Further, it is clear from the plot of the scores on the first two principal components (see Figure 5.17) that these do not fall into four distinct cluster. This reflects the fact that in the real world species' spatial behaviour does not fall into clear-cut categories but form a continuum. As a result, one

should mainly focus on the interpretation of the fist principal components as a characterisation of the community rather than on overinterpreting each species' position in the dendrogram in Figure 5.16.

	resprouter	seeder
group I	*Constylis juncea RH* *Hibbertia hypericoides SH* *Isopogon linearis SH* *Lepidosperma tenue SE* *Lomandra sp. SE* *Chordifex sinuosus RH* *Scholtzia aff. involucrata SH*	*Leucopogon constephioides SH*
group II	*Alexgeogea nitens RH* *Eremaea astrocarpa SH* *Hypocalymma xanthopetalum SH* *Lyginia barbata RH* *Melaleuca scabra SH***	*Andersonia heterophylla SH* *Astroloma xerophyllum SH* *Conospermum crassinervium SH* *Leucopogon striatus SH*
group III	*Banksia attenuata TSH* *Banksia menziesii TSH* *Conostylis canesence RH* *Bossia eriocarpa SH* *Eriostemon spicatus SH*** *Hibbertia crassifolia SH* *Hibbertia subvaginata SH* *Jacksonia floribunda SH* *Phebocarya filifolia RH*	*Boronia ramosa SH* *Stylidium crossocephalum STI*
group IV	*Conostephium pendulum SH* *Dasypogon bromeliifolius RH* *Patersonia occidentalis SE*	

Table 5.6: Species grouped by their spatial behaviour and fire regeneration method. ** Indicates that the species is both a seeder and a resprouter. growth habit: RH rhizomatous, SH shrub, SE sedge, TSH tall shrub, STI stilt plant

Chapter 6

Modelling

A fundamental problem in ecology is assessing the nature and strength of interaction between individuals from the same and from different species (Law et al. 1997). The previous chapters have mainly considered and characterised intra-specific interaction amongst individuals. It is only through addressing inter-specific interaction that we are able to assess fundamental questions about species coexistence and biodiversity. This includes for instance whether, as niche theory predicts, the identity of neighbours matters, i.e. whether individuals interact in the same way with all other individuals irrespective of their identity (Uriate et al. 2004).

However, interaction strength is not directly observable or measurable and needs to be operationalised through parameters in a model. Here, we take three approaches to fitting specific spatial point process models to the Cooljarloo data set investigating inter-specific interaction in particular. The parameters of the models reflect characteristics of the pattern such as its intensity as well as the strength and the nature of the interaction, i.e. whether it is positive or negative.

However, spatial point process models have not been fitted to entire plant communities before. Most applications have been either of a merely descriptive nature using Ripley's K-function (Haase 1995; Dale et al. 2002; Liebhold and Gurevitch 2002; Wiegand and Moloney 2004) or the model was restricted to a very small number of species, typically not more than two or three (Mateu et al. 1998). In addition, an increasing number of datasets detailing individuals in communities with a large number of different species are becoming available (see e.g. Burslem et al. 2001). Again, the problem faced here is the complexity of the data set. We have sought to apply parsimonious approaches such that parameter estimation is feasible with regard to running times and tractability despite the data's complexity.

We begin with a simple approach which assumes equal mutual interaction between individuals from different species within an interaction zone around each individual, the Continuum Ising model (see Section 6.1). As it turns out, this approach is too simplistic. As a consequence, we attempt to improve on the model by first introducing an asymmetric interaction structure (Section 6.2.3) using a frequentist approach. We eventually also vary the interaction radius within a Bayesian setting (Section 6.3). We assess the fit for each of the approaches and derive recommendations as to further approaches (Illian et al. 2006).

6.1 Continuum Ising model

Due to the complexity of the data set, the initial modelling attempt fits a spatial point process model with a simple structure to improve interpretabil-

ity. This is a model with as small a number of parameters as possible, in order to facilitate parameter estimation as well as interpretation of results. As a consequence, a multi-type Continuum Ising Model, as introduced in Section 3.4.4.3, is fitted, see also Møller and Waagepetersen (2003b). Furthermore, given the size of the data set and the potentially large number of parameters we aim here at finding an efficient estimation method. We show that an estimation procedure based on the pseudolikelihood, similar to the one introduced by Baddeley and Turner (2000), see Section 3.5.3, may be applied here as well.

This model assumes that, given the locations of the individuals from the other species, the pattern of each species follows an inhomogeneous Poisson process. This implies that any observed inhomogeneity, clustering or regularity is entirely due to the mutual interaction with other species, and not a result of environmental variation. This assumption might appear rather simplistic and unrealistic, but the assumption of environmental homogeneity holds for the given data set (Dixon 2005; Armstrong 2005).

The model also assumes that interactions between individuals from different species are symmetric, i.e. that there is mutual inhibition or attraction respectively of the same strength between individuals from the pairs of species. Figure 6.1 demonstrates the interaction structure for this model for the case of two species. The arrows indicate the direction of the interaction as well as its strength. Here, species 1 interacts with species 2 and species 2 interacts with species 1. The strength of both these interactions is the same, as indicated by the equally thick arrows. This model may be extended canonically

to include more than two species while still maintaining a pair-wise interaction structure. We will proceed by initially fitting a two-species model.

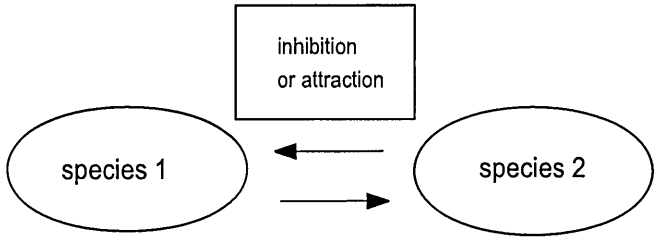

Figure 6.1: Interaction structure in the Continuum Ising model for two species

More technically, the Continuum Ising model may be defined through its density function, as explained in Section 3.4.4.3. Consider a multivariate spatial point process $X \in S$, consisting of k different types of points $1, \ldots, k$. I.e. we have k separate subprocesses for k different species, $X_1, \ldots, X_k$, with the

following joint probability density function with regard to the homogeneous Poisson process:

$$f(x_1, \ldots, x_k) = \frac{1}{c(\theta)} \prod_{i=1}^{k} \beta_i^{n(x_i)} \prod_{1 \le i < j \le k} \prod_{\zeta \in x_i} \prod_{\eta \in x_j} h_{ij}(\|\zeta - \eta\|). \tag{6.1}$$

Here, θ is a vector of parameters: $\frac{1}{c(\theta)}$ the normalising constant, β_i is a parameter governing the intensity of the process X_i, $h_{ij}(\|\zeta - \eta\|)$ is an interaction function describing the interaction between points from processes X_i and X_j, where $i \neq j$ and $n(x_i)$ denotes the number of points in process i.
We choose the following interaction function:

$$h_{ij}(\|\zeta - \eta\|) = \exp\left(\frac{\theta_{ij}}{\|\zeta - \eta\|}\right), \tag{6.2}$$

where $\|.\|$ denotes the absolute distance, or L_1-norm between two points and $\theta_{ij} = \theta_{ji}, \theta_{ij} < 0$.
Note that only the interaction between different types of points is modelled here, such that the conditional distribution of process x_i follows an inhomogeneous Poisson process, given the locations of the other processes. Also, with the interaction function in equation 6.2 the model only accounts for negative interaction, i.e. $\theta_{ij} < 0$. Allowing $-\infty < \theta_{ij} < \infty$ results in problems with integrability.

Figure 6.2 shows the interaction function for different interaction parameters on the unit square. The strength of interaction decreases with the distance from the individual plant. A larger interaction parameter leads to stronger inhibition, i.e. the probability for another individual occurring in its vicinity decreases. The interaction function gradually approaches one,

signifying no interaction. When the interaction parameter increases, the interaction function converges later.

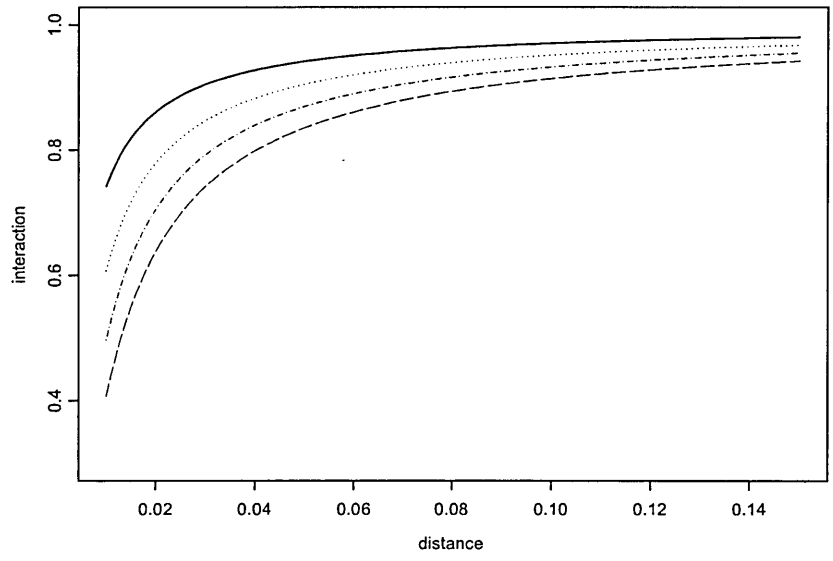

Figure 6.2: Interaction function for Continuum Ising model with different interaction parameters; full line: $\theta_{12} = -0.003$, dotted line: $\theta_{12} = -0.005$, dashed-dotted line: $\theta_{12} = -0.007$ and dashed line: $\theta_{12} = -0.009$.

6.1.1 Model fitting

We aim at applying a parsimonious modelling approach and hence aim at finding an equally parsimonious estimation procedure which yields reliable results. We thus follow the approach of Baddeley and Turner (2000) in favour of applying established software for estimation using a pseudolikelihood approach. The following section derives the approximate estimate of

the log-pseudolikelihood based on a generalised linear model using readily available software.

6.1.1.1 Maximum pseudolikelihood

In Section 3.5.3 we described an approach that may be used for parameter estimation when the normalising constant is analytically intractable. Here, the normalising constant $c(\theta)$ is likewise not known analytically. Since the above density is an exponential family density, we can choose maximum pseudolikelihood approaches for parameter estimation as described in Section 3.5.3 above.

The pseudolikelihood for parameter vector $\theta = (\beta_i, \theta_{ij})$ can be written as

$$\mathrm{PL}(\theta) = \prod_{i=1}^{k} f(x_i | x_{-i}),$$

where x_{-i} denotes all processes $x_1, \ldots, x_k$ other than x_i and $f(.)$ is the density function of an inhomogeneous Poisson process with intensity function

$$\xi \to \beta_i \prod_{1 \le i < j \le k} \prod_{\eta \in X_j} \exp\left(\frac{\theta_{ij}}{\|\xi - \eta\|}\right).$$

Hence, the pseudolikelihood for the model in (6.1) is

$$\begin{aligned} PL(\theta) = \quad & \prod_i \prod_{\xi \in X_i} \exp\left(\ln \beta_i + 2 \sum_{j>i} \theta_{ij} \sum_{\xi \in X_i} \sum_{\eta \in X_j} \frac{1}{\|\xi - \eta\|}\right) \cdot \\ & \exp\left[-\sum_i \int_W \exp\left(\ln \beta_i + 2 \sum_{j>i} \theta_{ij} \sum_{\eta \in X_j} \frac{1}{\|\xi - \eta\|}\right) d\xi\right] \end{aligned}$$

This is an exponential family model and we now consider the following approximation using a quadrature scheme:

$$\sum_i \int_W \exp\left(\ln\beta_i + 2\sum_{j>i}\theta_{ij}\sum_{\eta\in X_j}\frac{1}{\|\xi-\eta\|}\right)d\xi$$
$$\approx \sum_i \sum_{\text{quad}} \exp\left(\ln\beta_i + 2\sum_{j>i}\theta_{ij}\sum_{\eta\in X_j}\frac{1}{\|\xi-\eta\|}\right)w_\xi,$$

where the w_ξ are quadrature weights with $w_\xi = \frac{|C_h|}{\sum_i(1+n(x_i\cap C_h))}$ if $\xi \in C_h$ and $\sum w_\xi = |W|$.

We can thus approximate the log-pseudolikelihood through

$$\log PL(\theta) \approx \sum_i \sum_\xi (y_{i\xi}\log\lambda^*_{i\xi} - \lambda^*_{i\xi})w_\xi \tag{6.3}$$

with

$$y_{i\xi} = \frac{\mathbf{1}[\xi \notin C]}{w_\xi}$$

and

$$\lambda^*_{i\xi} = \exp(\log\beta_i + 2\sum_{j,j>i}\theta_{ij}\sum_{\eta\in X_j} 1/\|\xi-\eta\|).$$

The right-hand side of (6.3) is formally equivalent to the weighted log likelihood of a generalised linear model with Poisson responses. Baddeley and Turner (2000) suggest using algorithms provided by software packages such as `S-PLUS` or `R` to estimate the parameters of the generalised linear model (see also Section 3.5.3 and Baddeley and Turner (2005)). This approach is particularly useful in the given situation where the complexity of the data structure makes the estimation process susceptible to errors, as it greatly facilitates and stabilises the estimation process.

For the model described in (6.1) we can thus relate the conditional intensity function to the vector of sufficient statistics t as follows

$$g(\lambda_\theta(u;x)) = \theta^T \cdot t,$$

where g is a link function and $\theta = (\theta_{ij})$, where $\theta_{ii} = \ln \beta_i$ is the parameter vector and $t = (t_{ij})$ with

$$(t_{ii})_{i'\xi} = \mathbf{1}\{i = i'\}, \qquad \text{for } \xi \in X_{i'}$$

and

$$(t_{ij})_{i'\xi} = \mathbf{1}\{i = i'\} \cdot 2 \sum_{\eta \in X_j} 1/|\xi - \eta|.$$

The vectors t_{ii} and t_{ij} serve as the explanatory variables in the sense of glm. To estimate the model parameters we can use the following call to the generalised linear model function in S-PLUS to estimate the parameters

```
glm(y ~ t12+t13+...+tij+...+tkk, family=poisson, link=log,
weights=w)
```

with $1 \leq i \leq j \leq k$.

6.1.1.2 Estimated parameters for pairs of species

Initially, the model was fitted to the patterns formed by a subset of pairs of species in order to assess the suitability of the approach and to reveal potential weaknesses. The subset comprised those pairs where biological information existed that suggested negative inter-specific interaction (Armstrong 1991, 2005; Dixon 2005). The parameters for pairs in this subset were

estimated as described above. Initial focus has been on the resprouting species *Scholtzia involucrata* (species number 61), which was assumed to repulse most other species as it has been established in the same location for a very long time. In Chapter 4 we found an indication of negative interaction for the species with *Leucopogon striatus* (48) but no interaction with *Lyginia barbata* (50) using the two-dimensional L-functions.

A permutation test (Good 2000) was developed and applied to test for significant interaction parameters using random labelling. In each iteration, the locations of the individual plants were fixed but the labels (species one and species two, respectively) were randomly allocated to these locations. For each permutation, the parameters were estimated in the same way as for the original labelling. Interaction parameters were considered significant when the estimated parameter for the original labelling was larger than 95% of the estimated interaction parameters for the permuted patterns. Table 6.1 lists some pairs of species with estimated interaction parameters and significance level of the permutation test.

Apparently, *Scholtzia involucrata* interacts negatively with a number of species in the data set, but there are also species that the resprouter does not interact with. This is in line with the results found in the exploratory data analysis.

Figure 6.3 shows the graphs of the interaction function for *Dampiera linearis* (22, full line) and *Scholtzia involucrata* (61) as well as *Leucopogon striatus* (48) and *Scholtzia involucrata* (dashed line). The dashed line approaches the value 1 more rapidly, indicating that *Scholtzia involucrata* repulses *Leucopogon striatus* less and within a smaller range around the individuals than

species numbers	interaction parameter θ_{ij}	significance level
48 and 61	-2.511	$\mathbf{p = 0.030}$
22 and 61	-4.486	$\mathbf{p = 0.000}$
22 and 48	-0.016	$p = 0.857$
22 and 36	-0.4475	$p = 0.732$
23 and 61	-2.963	$\mathbf{p = 0.000}$
25 and 61	-1.352	$\mathbf{p = 0.000}$
38 and 61	-1.321	$\mathbf{p = 0.000}$
50 and 61	-0.08903	$p = 0.643$
59 and 61	-0.01915	$p = 0.0870$

Table 6.1: Pairs of species with estimated interaction parameters and significance levels; species numbers are: *Dampiera linearis* (22), *Dasypogon bromeliifolius* (23), *Eremaea astrocarpa* (25), *Hypocalymma xantopethalum* (36), *Jacksonia floribunda* (38), *Leucopogon striatus* (48), *Lyginia barbata* (50), *Platysace juncea* (59) and *Scholtzia involucrata* (61); significant parameters in bold

it repulses *Dampiera linearis.*

6.1.2 Model validation

In order to validate the model, i.e. to assess its fit to the data, inhomogeneous L-functions were calculated for the pattern of one species as well as for 39 simulations from the model for that species given the respective other species as described in Møller and Waagepetersen (2003b). Taking the maximum and minimum of these at each distance, simulation envelopes were constructed. Figure 6.4 shows the plot of the inhomogeneous L-function for the species *Dampiera linearis* given *Scholtzia involucrata* with simulation envelopes of 39 bivariate Continuum Ising point processes with the estimated parameters as above. Figure 6.5 shows the similar plot for the species *Scholtzia involucrata*

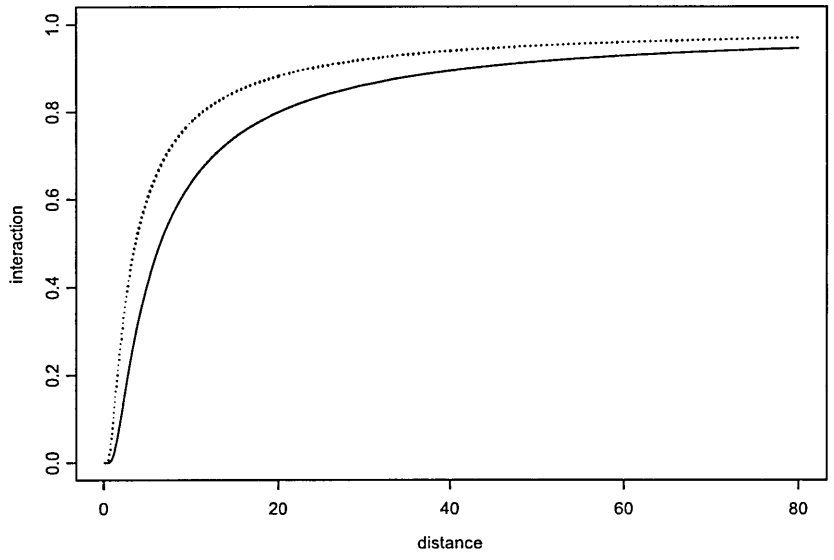

Figure 6.3: Interaction functions for Continuum Ising model for species *Dampiera linearis*(22) and *Scholtzia involucrata* (61) with $\phi_{ij} = -2.511$ and $p = 0.030$ (full line) and *Leucopogon striatus* (48) and *Scholtzia involucrata* with $\phi = -4.486$ and $p = 0.000$ (dashed line)

given *Dampiera linearis.* The model explains the locations of the plants quite well since the estimated functions both remain inside the envelopes.

However, Figure 6.6 shows the inhomogeneous L-function for the species *Leucopogon striatus* given species *Scholtzia involucrata* with simulation envelopes of 39 bivariate point patterns generated from the Continuum Ising model with parameters as above. The model fit for *Leucopogon striatus* is clearly not as good as in the previous case, as the L-function is outside the envelopes indicating clustering that has not been explained by the model.

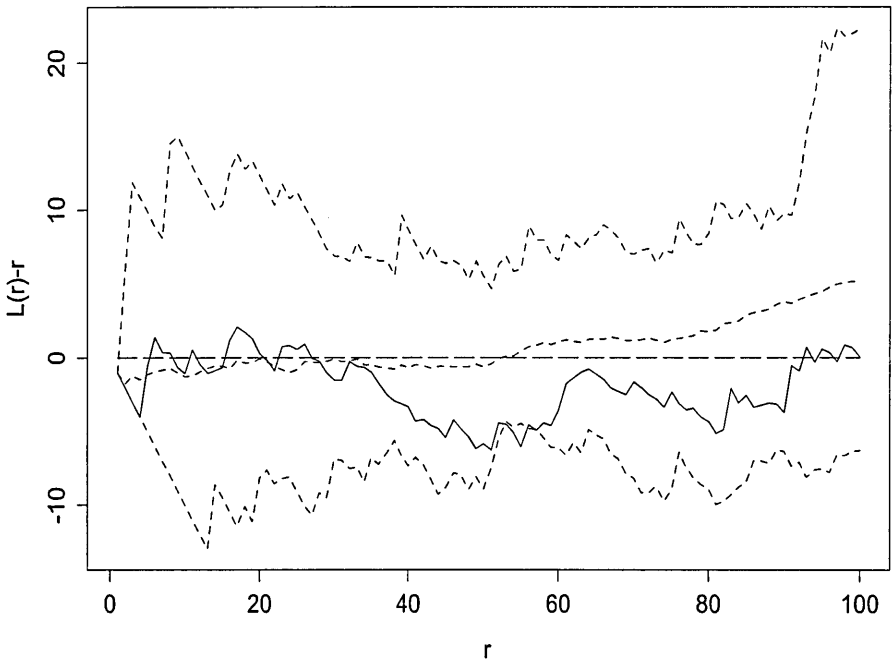

Figure 6.4: Inhomogeneous L-function for species *Dampiera linearis* given the species *Scholtzia involucrata* with simulation envelopes for the Continuum Ising model. The dashed lines are the upper simulation envelope, mean of the L-functions at each distance and lower simulation envelope respectively. The full line is the estimated L-function for the data.

6.1.3 Conclusions

As we can see from the plots of the inhomogeneous L-functions, the model fits the data reasonably well in some cases. For instance, the resprouting species *Scholtzia involucrata* (61) given the locations of species *Dampiera linearis* (22). However, the above plots clearly indicate that the model does not fit the data in all cases. In particular, for the seeder species *Leucopogon striatus* (48) we find that not all of the aggregation can be explained by the interaction with species *Scholtzia involucrata* (61).

The model assumes a symmetric interaction structure where two species show

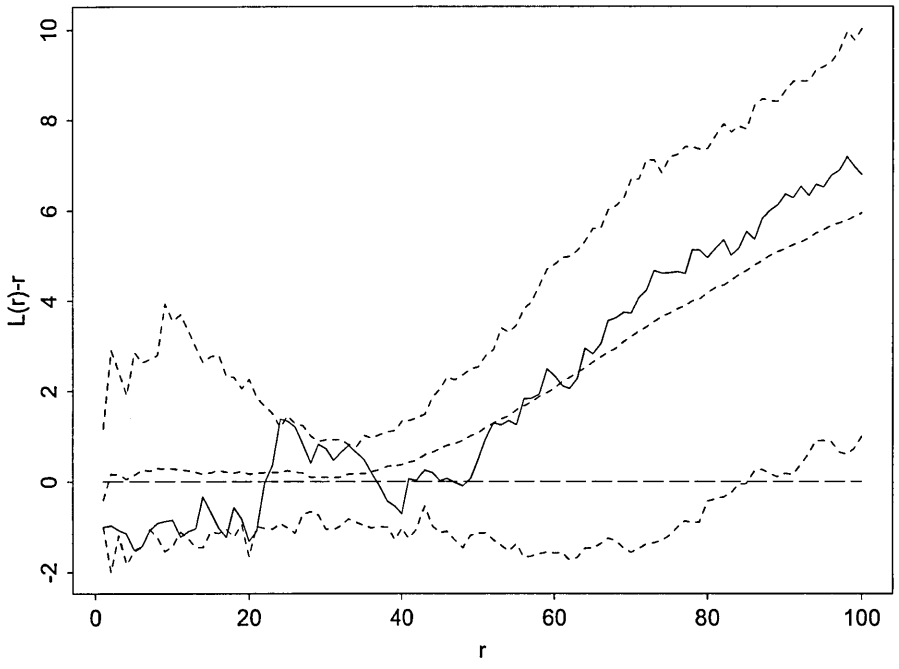

Figure 6.5: Inhomogeneous L-function for species *Scholtzia involucrata* given the species *Dampiera linearis* with simulation envelopes for the Continuum Ising model. The dashed lines are the upper simulation envelope, mean of the L-functions at each distance and lower simulation envelope respectively. The full line is the estimated L-function for the data.

mutual repulsion. In general, this assumption is not very realistic in an ecological context. Typically, one species would, say, inhibit the growth of the other much more than vice versa since competition tends to be asymmetric (Watkinson 1997). Considering the example of the two species *Leucopogon striatus* (48) and *Scholtzia involucrata* (61) we can clearly see that the model is too simple. This might be due to the assumption of mutually equal interaction not being valid in this case. In fact, *Scholtzia involucrata* (61) is a resprouter and *Leucopogon striatus* (48) is a seeder; we cannot assume that the resprouters are inhibited by the seeders in the same way as the seeders

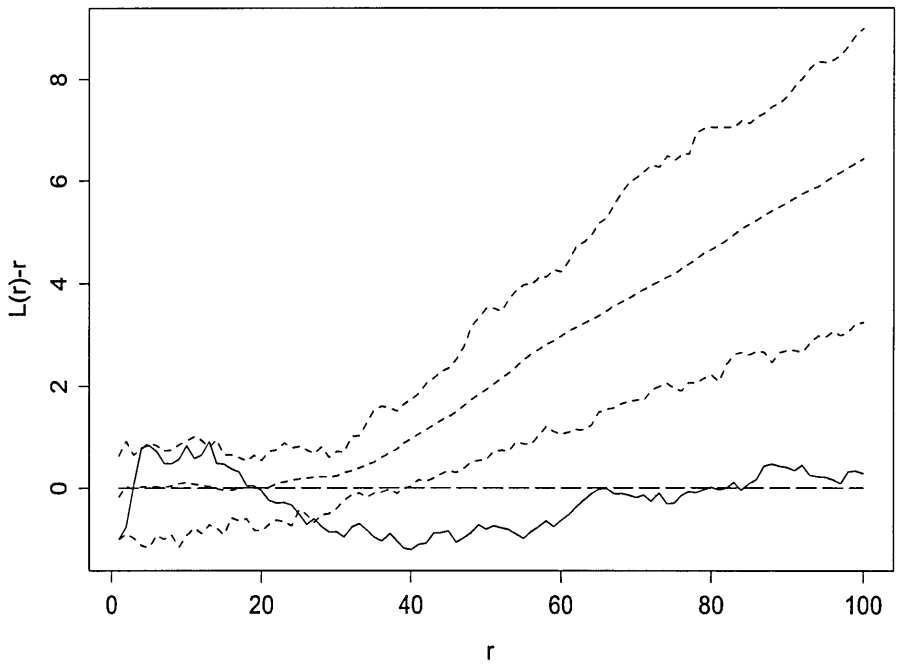

Figure 6.6: Inhomogeneous L-function for species *Leucopogon striatus* given *Scholtzia involucrata* (61) with simulation envelopes for the Continuum Ising model. The dashed lines are the upper simulation envelope, mean of the L-functions at each distance and lower simulation envelope respectively. The full line is the estimated L-function for the data.

are inhibited by the resprouters. The resprouters may have established themselves at their current locations for possibly hundreds of years whereas the seeders have started as young plants after the last fire, i.e. at most ten years before the data was collected (Armstrong 2005). This apparent asymmetry is not captured by the model.

This aspect has been identified as the main weaknesses of the above approach. Due to the apparent inappropriateness of the model we will not pursue fitting the model to more than two species here. We still believe that the Continuum Ising model may have applications outwith the context

of this dataset, and thus feel that the simplified estimation procedure may be usefully applied in many other contexts where a symmetric interaction structure is a realistic assumption.

The following sections describe a more refined approach which accounts for the asymmetric interaction, using a hierarchical model which will resolve the shortcoming described above. It will take the different fire regeneration strategies of the species into account by modelling the seeding species given the location of the resprouting species. It will also use a different interaction function that allows both negative and positive inter-species interaction.

6.2 Hierarchical model, frequentist approach

Since the resprouting plants have been at exactly the same location for a very long time (some of them for hundreds and even thousands of years (Armstrong 1991; Dixon 2005)), the seeders, which start anew after each fire, do so with the resprouters already present. We thus assume that the growth of seeder plants will be influenced by the resprouting plants already established in the plot, whereas an influence of the seeders on the resprouting plants is highly unlikely. Figure 6.7 illustrates the interaction structure for two species which may be canonically generalised to more species as described below.

In order to capture the asymmetric interaction, the model uses a hierarchical structure assuming that the locations of the resprouting plants are independent of those of the seeding plants, whereas the seeding plants are modelled conditionally on the locations of the resprouting plants.

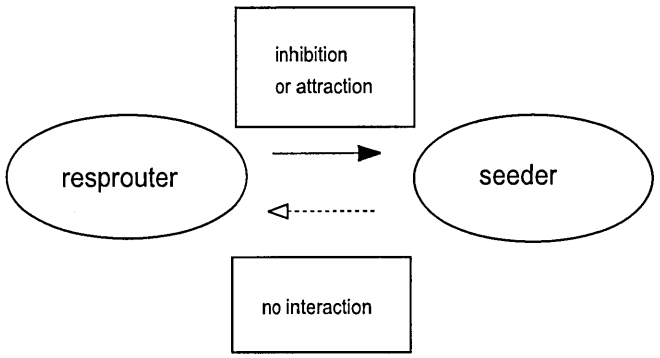

Figure 6.7: Asymmetric interaction structure in the hierarchical model for two species

6.2.1 The model

In the following, we consider the 24 point patterns in Figure 6.8, i.e. the 5 most abundant species of seeders and the 19 species of resprouters suspected to be most influential either as a result of their abundance (resprouters 1-13) or their size (resprouters 14-19) (Armstrong 2005).

We assume that each individual resprouter plant has a circular zone of influence with the strength of influence decreasing with the distance from each plant. Beyond this zone the influence is considered zero. The interaction

Table 6.2: Range of zone of influence (in cm) for the resprouters considered in the hierarchical model

1. *Alexgeorgia nitens*:	10-40	11. *Phlebocaria filifolia*:	20-30
2. *Conostylis canescens*:	5-15	12. *Chordifex sinuosus*:	25-75
3. *Dasypogon bromeliifolius*:	15-60	13. *Scholtzia involucrata:*	30-50
4. *Eremae astrocarpa*:	25-75	14. *Allocasuarina humilis*:	50-130
5. *Hibbertia crassifolia*:	10-25	15. *Banksia attenuata*:	150-400
6. *Hibbertia hypericoides*:	10-20	16. *Banksia grandis*:	50-200
7. *Hibbertia subvaginata*:	10-25	17. *Banksia ilicifolia*:	50-200
8. *Hypocalymma xantopetalum*:	10-25	18. *Banksia menziesii*:	50-250
9. *Lomandra sp.*:	2-10	19. *Eucalyptus todtiana*:	10-250
10. *Lyginia barbata*:	20-100		

may be positive (e.g. caused by subcanopy soil enrichment (Callaway 1995)) or negative (e.g. caused by allelopathy (Armstrong 1991)). Prior information (Armstrong 2005) available on the typical radii of the influence zones is given in Table 6.2 which also shows the numbering we use for resprouter species. The five species of seeders are 1. *Andersonia heterophylla*, 2. *Astroloma xerophyllum*, 3. *Conospermum crassinervium*, 4. *Leucopogon conostephioides* and 5. *Leucopogon striatus*.

6.2.2 Likelihood

We use the following statistical model.

Denote by W the 22 m by 22 m plot where the plants were recorded, $x_1, \ldots, x_{19}$ the observed point patterns for the 19 resprouters, $y_1, \ldots, y_5$ the observed point patterns for the 5 seeders, and $X_1, \ldots, X_{19}, Y_1, \ldots, Y_5$ the corresponding spatial point processes, i.e. here each X_j or Y_i is considered to be a random finite subset of W. Since we are mainly interested in the inter-species

interactions between seeders and resprouters, we leave the marginal distributions of the resprouters unspecified and restrict attention to the conditional likelihood of the seeders given the resprouters.

Conditional on $X_1 = x_1, \ldots, X_{19} = x_{19}$ we assume that $Y_1, \ldots, Y_5$ are independent Poisson processes, i.e. that there is no intra-species interaction for the seeder species. We also assume that each Y_i has intensity function

$$\lambda(\xi|x, \theta_i) = \exp\left(\theta_i s(\xi|x)^{\mathsf{T}}\right), \qquad \xi \in W, \tag{6.4}$$

where we use the following notation: $x = (x_1, \ldots, x_{19})$ is the collection of all 19 resprouter point patterns; $\theta_i = (\theta_{i0}, \ldots, \theta_{i19})$ is a vector of parameters, where $\theta_{i0} \in \mathbb{R}$ is an intercept and for $j = 1, \ldots, 19$, $\theta_{ij} \in \mathbb{R}$ controls the influence of the jth resprouter on the ith seeder (a positive value of θ_{ij} means a positive/attractive association; a negative value of θ_{ij} means a negative/repulsive association); $s(\xi|x) = (1, s_1(\xi|x), \ldots, s_{19}(\xi|x))$ with

$$s_j(\xi|x) = \sum_{\eta \in x_j} h_\eta(\|\xi - \eta\|), \qquad j = 1, \ldots, 19,$$

where $\|\cdot\|$ denotes Euclidean distance; and h_η is a smooth interaction function given by

$$h_\eta(r) = \begin{cases} (1 - (r/R_\eta))^2)^2 & \text{if } 0 < r \leq R_\eta \\ 0 & \text{else} \end{cases}$$

for $r \geq 0$, where $R_\eta \geq 0$ defines the radii of interaction of a given resprouter at location η, cf. Table 6.2.

Thus, given the resprouters x the number N_i of points in Y_i is Poisson distributed with mean value $\int_W \lambda(\xi|x, \theta_i)\mathrm{d}\xi$, and if we also condition on N_i, the points in Y_i are independent and identically distributed with a density

proportional to $\lambda(\xi|x,\theta_i)$. It follows that the log likelihood function based on the 5 seeder point patterns $y = (y_1, \ldots, y_5)$ is

$$l(\theta, R; y|x) = \sum_{i=1}^{5} \Big[\theta_i \sum_{\xi \in y_i} s(\xi|x)^{\mathsf{T}} - \int_W \exp(\theta_i s\,(\xi|x)^{\mathsf{T}}))\,\mathrm{d}\xi\Big], \tag{6.5}$$

where $\theta = (\theta_1, \ldots, \theta_5)$ is the vector of all 100 parameters θ_{ij} and R is the vector of all 3168 radii R_η, $\eta \in x_j$, $j = 1, \ldots, 19$ of the 3168 resprouter plants from the 19 species. In comparison, there are $N_1 + \cdots + N_5 = 1954$ seeder plants.

A potential weakness of our model is the assumed conditional independence between the seeders, see Sections 6.2.5 and 6.3.3 on model assessment.

6.2.3 Maximum likelihood estimation

For a fixed value of $\mathbf{R}$, due to the log linear intensity (6.4), the log likelihood (6.5) can be maximized relatively easily with respect to θ using the Berman and Turner (1992) device on the likelihood, similar to the estimation method introduced in Section 3.5.3 for the pseudolikelihood. In essence, the log likelihood is approximated by a weighted log likelihood of independent Poisson variables which can be optimised using standard software, such as the function `glm` in `R` or `SPLUS`, see also the `R` package `spatstat` (Baddeley and Turner 2005). In a frequentist setting, maximisation with respect to $\mathbf{R}$ on the other hand is difficult due to the high dimensionality of $\mathbf{R}$ and since the R_η-values do not enter the likelihood function in a log linear fashion. In this section we therefore make the simplifying assumption that for each resprouter type j, all R_η with $\eta \in \mathbf{x}_j$ are equal to a common interaction

radius R_j, given by the midpoint of the interval for the zone of influence in Table 6.2 (Armstrong 2005).

Given the chosen value of $\mathbf{R}$ another problem is whether the maximum likelihood estimate (MLE) of θ exists. By exponential family results (Barndorff-Nielsen 1978), the MLE for θ_i exists if and only if $t_j(\mathbf{y}_i) \equiv \sum_{\xi \in \mathbf{y}_i} s_j(\xi|\mathbf{x}) > 0$, $j = 1, \ldots, 19$. Depending on the value of R_j it may well happen that $t_j(\mathbf{y}_i) = 0$, and indeed this is sometimes the case with our choice of the R_j. In Figure 6.9, the fields marked with NA correspond to the 18 parameters where the MLE does not exist. One may choose to let $\theta_{ij} = -\infty$ if $t_j(\mathbf{y}_i) = 0$. This corresponds to a hard core effect of the jth resprouter on the ith seeder, cf. resprouter 2, 9, and 14. It may on the other hand also simply be the case that the chosen interaction radii for these resprouters are too small.

6.2.4 Results

The ML estimation resulted in negative as well as positive interaction parameters, signifying repulsion and attraction, respectively. See Table 6.3 for the interaction parameters. A $*$ indicates a parameter that has been considered significant, i.e. a parameter where the approximate confidence intervals derived from the standard errors returned from the glm routine do not cover zero.

To summarize this, Table 6.4 lists all significant positive and negative interactions for all resprouter species. There were clearly more negative than positive interactions. To provide a clearer overview we consider z-statistics

Table 6.3: Estimated interaction parameters for the frequentist approach

	seeder 1	seeder 2	seeder 3	seeder 4	seeder 5
θ_1	-0.293	-0.818	-0.697^*	-0.361^*	-0.955^*
θ_2	NA	NA	NA	NA	NA
θ_3	-0.654^*	0.337	-0.198	-0.213	-0.383
θ_4	-0.446^*	0.001	-0.084	-0.470^*	-0.034
θ_5	-1.457	-0.700	-8.143	1.708	-0.161
θ_6	0.059	-1.307	-0.286	-0.489	-3.955
θ_7	-0.163	-4.644	-0.256	0.399	-2.198
θ_8	-1.575	NA	-2.363	-1.347	0.875
θ_9	NA	NA	NA	NA	NA
θ_{10}	-0.153	0.433	-0.109	-0.222	0.161
θ_{11}	-0.583	0.873	-1.257	-0.092	-0.255
θ_{12}	0.170	-0.260	-0.368	0.247	0.225
θ_{13}	-0.173	-1.387	-0.034	-0.526^*	-0.959^*
θ_{14}	-0.293	NA	NA	NA	NA
θ_{15}	0.087	0.768^*	0.038	0.095	-0.272^*
θ_{16}	0.744	NA	-1.669	0.115	0.991
θ_{17}	-0.477	NA	1.517	-4.631	2.145
θ_{18}	0.102	0.740^*	0.192	0.219	-0.130
θ_{19}	NA	NA	-5.791	0.019	-1.434

Table 6.4: List of all significantly positive and significantly negative interactions for all resprouters, maximum likelihood estimation with constant interaction radii across individuals for each species

	positive interaction	negative interaction
1. *Alexgeorgia nitens*:	none	seeder 3, 4 and 5
2. *Conostylis canescens*:	none	none
3. *Dasypogon bromeliifolius*:	none	seeder 1
4. *Eremae astrocarpa*:	none	seeder 1 and 4
5. *Hibbertia crassifolia*:	none	none
6. *Hibbertia hypericoides*:	none	none
7. *Hibbertia subvaginata*:	none	none
8. *Hypocalymma xant.*:	none	none
9. *Lomandra sp.*:	none	none
10. *Lyginia barbata*:	none	none
11. *Phlebocaria filifolia*:	none	none
12. *Chordifex sinuosus*:	none	none
13. *Scholtzia involucrata*:	none	seeder 4 and 5
14. *Allocasuarina humilis*:	none	none
15. *Banksia attenuata*:	seeder 2	seeder 5
16. *Banksia grandis*:	none	none
17. *Banksia ilicifolia*:	none	none
18. *Banksia menziesii*:	seeder 2	none
19. *Eucalyptus todtiana*:	none	none

for the various θ_{ij}, i.e. MLEs scaled by the associated estimated standard errors. The upper plot in Figure 6.9 shows transformed z-statistics $\Phi(z)$, where Φ is the standard normal distribution function. Dark shading corresponds to large $\Phi(z)$-values. Assuming that the z-statistics are approximately standard normal (and hence that the $\Phi(z)$ are approximately uniform on [0,1]), the starred fields correspond to parameters which are significantly different from zero at the 5 % level. However, with 95 interaction parameters there is obviously an issue of multiple testing and hence we should be careful when interpreting the "starred" parameters. Rather than looking at individual parameters it seems more appropriate to look for resprouters or seeders with a consistent pattern of dark or light fields in Figure 6.9.

For example, it seems reasonable to conclude that resprouters 1 and 4 have a repulsive effect on the seeders whereas there may be evidence that resprouter 15 and 18 have an attractive effect on the seeders. So apparently the seeders 1, 3, 4 and 5 are repulsed by resprouting plants whereas seeder 2 is attracted by some of the resprouting plants.

6.2.5 Model validation

In order to check whether we have obtained reasonable MLE estimates we compared the integral of the estimated intensity surfaces for each of the seeders with its abundance and this resulted in approximately the same numbers in all five cases. The exact values of both the number of points in the seeder patterns and the integral of the estimated intensity surface are summarised in Table 6.5. The estimated values are only slightly above the original values.

Table 6.5: Comparison of observed and estimated intensity

seeder	number of plants	integral of estimated surface
1	689	699.27
2	91	91.26
3	266	267.58
4	657	666.39
5	251	252.46

We then assess the fitted model using L-functions, $L(r) = \sqrt{K(r)/\pi}$, where we obtain an estimate $\hat{L}(r; \mathbf{y}_i, \hat{\theta}_i)$ for the ith seeder by replacing K with the estimate

$$\hat{K}(r; \mathbf{y}_i, \hat{\theta}_i) = \sum_{\xi,\eta \in \mathbf{y}_i} \frac{\mathbf{1}[0 < \|\xi - \eta\| \leq r]}{\lambda(\xi|\mathbf{x}, \hat{\theta}_i)\lambda(\eta|\mathbf{x}, \hat{\theta}_i)} e_{\xi,\eta}$$

where $\mathbf{1}[\cdot]$ is the indicator function and $e_{\xi,\eta}$ is an edge correction, see Baddeley et al. (2000) and Møller and Waagepetersen (2003b). The estimated L-functions with 95% envelopes obtained by simulation under the fitted model (Møller and Waagepetersen 2003b) are shown in Figure 6.10. For seeder 1 the estimated L-function is clearly above the envelopes for distances up to about 400 cm, indicating that there is clustering present in the point pattern that is not explained by the model. This clustering might be a result of a mother-daughter relationship due to rhyzomatous growth behaviour or a sexual reproduction mechanism where offspring are located in the vicinity of the parent plant (Armstrong 2005). For seeders 2 and 5 it is not clear how to interpret the estimated L-functions, which are below the envelopes at larger distances. The model yields a reasonably good fit for seeders 3 and 4, though at very small distances there appears to be some intra-specific

repulsion. This might be the effect of a hardcore zone around each individual in which no conspecific individuals can survive.

The results indicate that a more appropriate model would have to take intra-specific interaction into account. Assuming known interaction radii, Högmander and Särkkä (1999) consider a hierarchical model with interactions for a multitype point pattern of ants' nests. However, in our situation, we believe that the assumption of known and equal interaction radii for resprouters of the same type is highly unrealistic since the plants vary in size and we choose to focus on this problem in the next section.

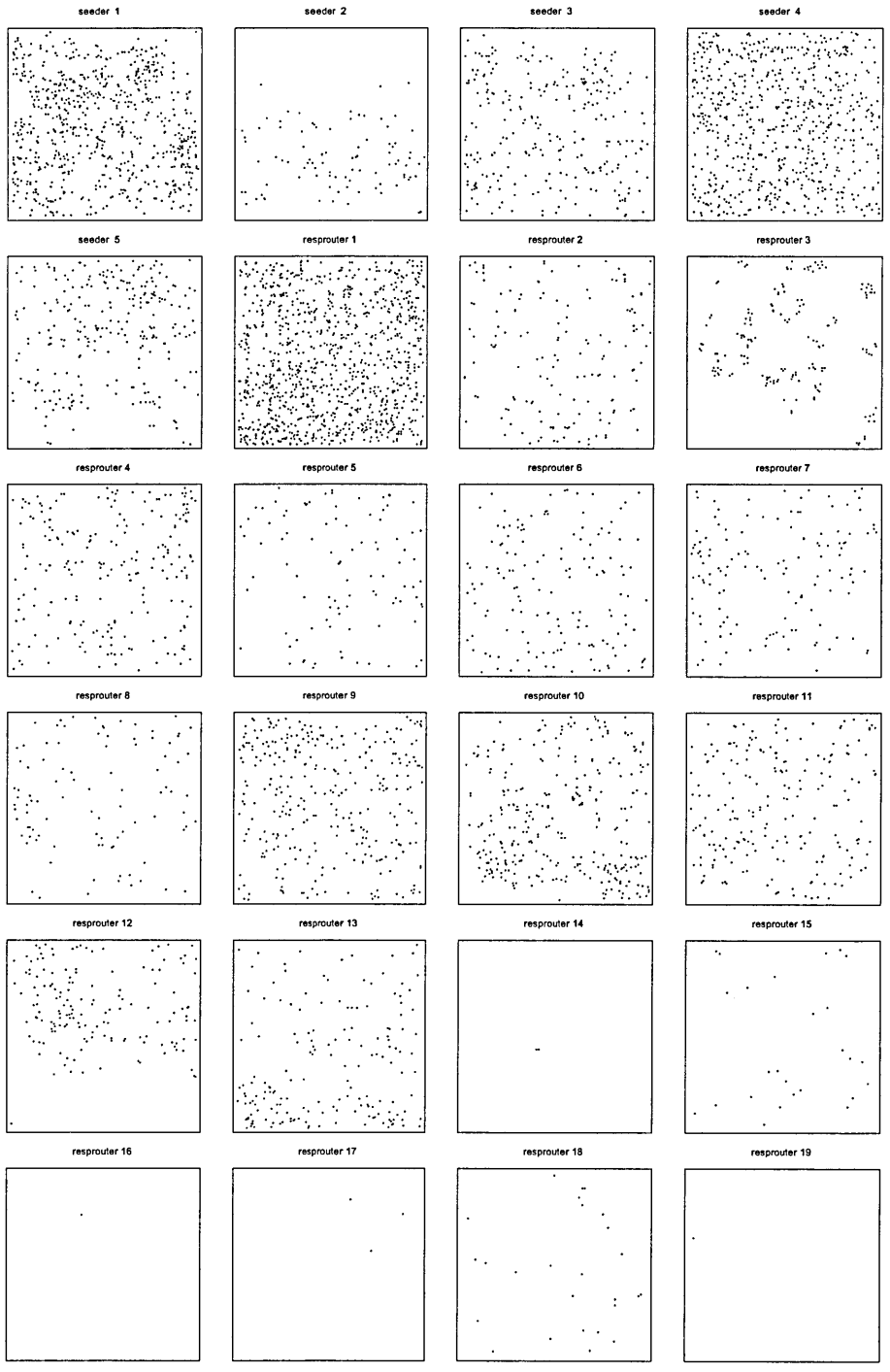

Figure 6.8: Observed point patterns for the 5 most abundant species of seeders and the 19 most influential species of resprouters.

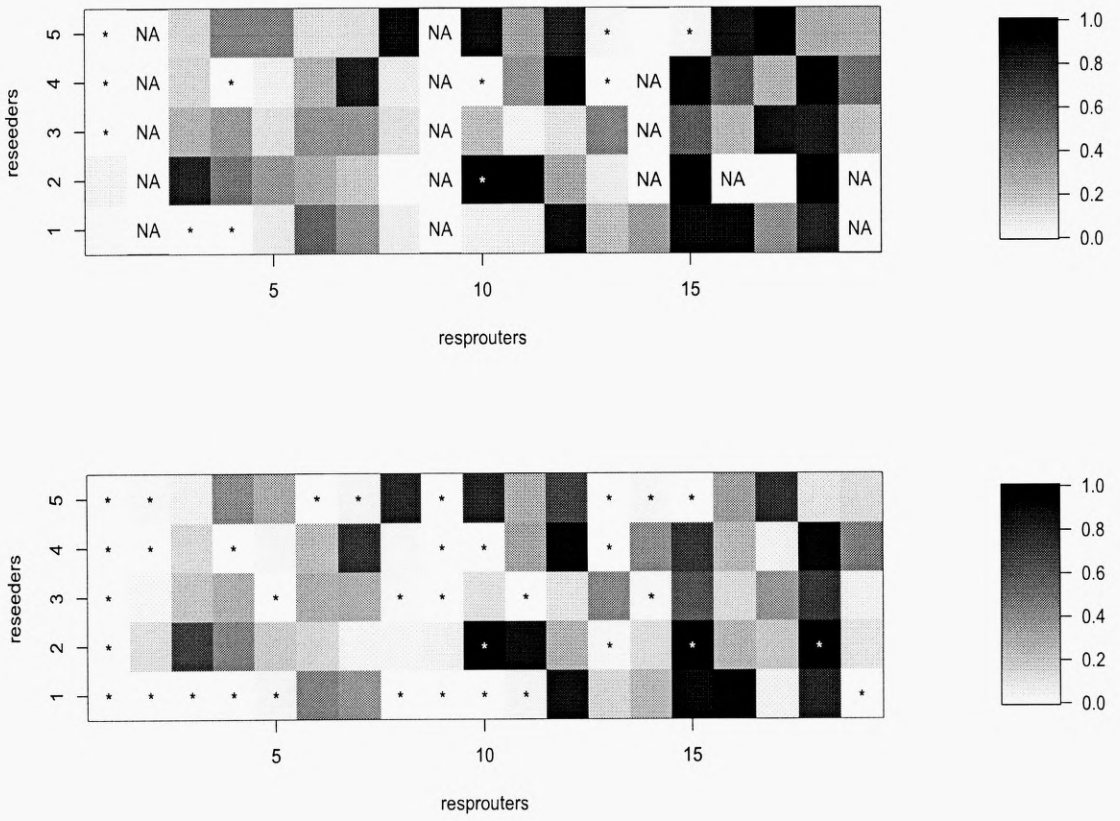

Figure 6.9: Upper plot: Grey scale plot of z-statistics for the interaction parameters θ_{ij}. Here the values are displayed as $u = \Phi(z)$, where Φ is the standard normal distribution function. Dark grey fields represent large values. Fields for parameters which are significantly different from zero at the 5 % level are marked with a *. Fields marked with NA correspond to θ_{ij} where the MLE does not exist. Lower plot: Grey scale plot of posterior probabilities $P(\theta_{ij} > 0|y)$ where dark grey fields indicate large values. The starred fields are those for which 0 is outside the central 95 % posterior interval for θ_{ij}.

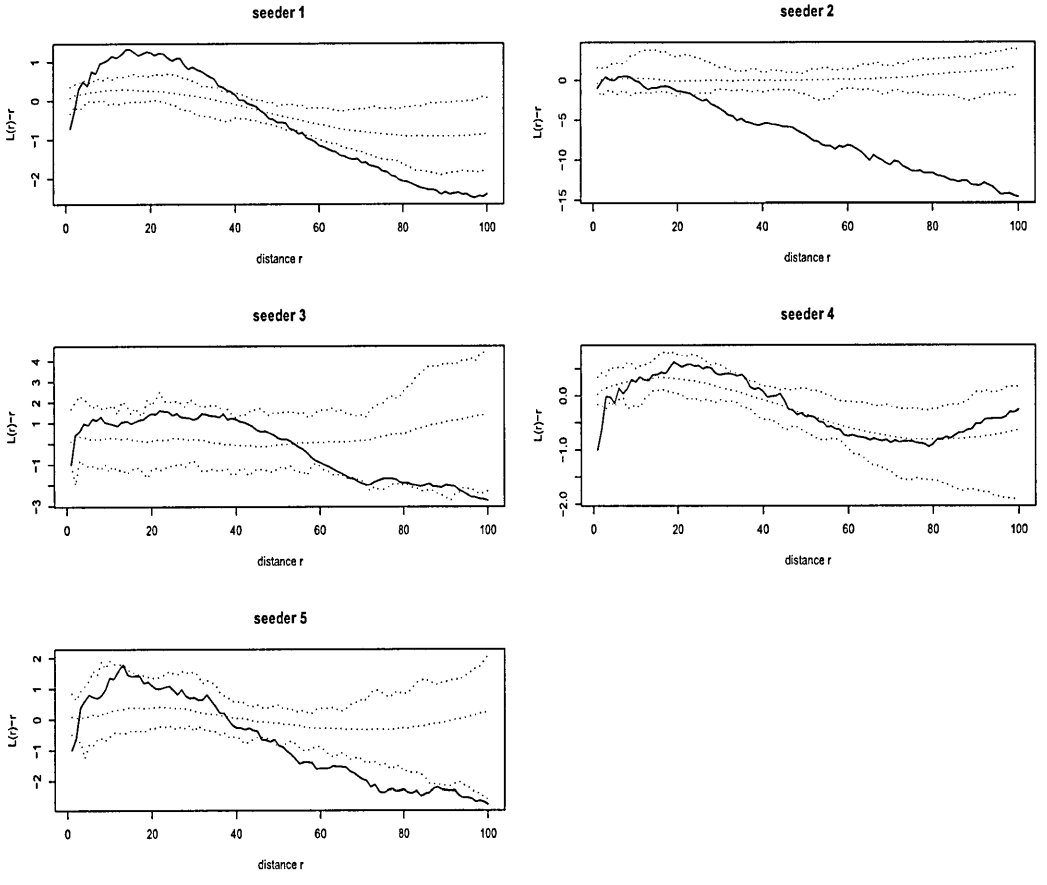

Figure 6.10: Estimated inhomogeneous $L(r) - r$-functions for seeder 1-5 with 95% envelopes simulated from the model. Distance r is in cm.

6.3 Hierarchical model, Bayesian approach

The assumption that all resprouters of the jth type share the same interaction radius R_j seems unrealistic but was needed in Section 6.2.3 for computational reasons. In this section we consider instead a more flexible Bayesian approach which has the potential to be applied to realistic yet complex data sets (Best et al. 1996). It allows us to assume different interaction radii for each individual resprouter plant. If data on the size of the species were available the interaction radius could be modelled as a function of this. This information is not available here, so we assume that a different interaction radius has to be estimated for each individual purely on the basis of the spatial pattern, i.e. the individuals' distance from the other plants. In addition, the approach allows us to incorporate prior information into the estimation, if available. Here, the prior distributions are based on the information in Table 6.2.

6.3.1 Prior and posterior

We make the following prior assumptions. The θ_{ij} and the R_η are independent random variables; each θ_{ij} follows a $\mathrm{N}(0, \sigma^2)$-distribution; for each $\eta \in x_j$, R_η follows the restriction of $\mathrm{N}(\mu_j, \sigma_j^2)$ to $[0, \infty)$, where (μ_j, σ_j^2) is chosen so that under the unrestricted $\mathrm{N}(\mu_j, \sigma_j^2)$ the range of the zone of influence in Table 6.2 is a central 95% interval; $(\theta, \mathbf{R})$ and X are independent, i.e. the posterior density for $(\theta, \mathbf{R})$ satisfies

$$\pi(\theta, \mathbf{R}|x, y) \propto \pi(\theta, \mathbf{R}) L(\theta, \mathbf{R}; \mathbf{y}|x).$$

Combining the prior assumptions with the log likelihood (6.5) we obtain the posterior density

$$\begin{aligned}
&\pi(\theta, \mathbf{R}|x, y) \\
&\propto \exp\Big(-\sum_{i=1}^{5} \Big[\theta_{i0}^2/(2\sigma^2) - \sum_{j=1}^{19} \Big\{\theta_{ij}^2/(2\sigma^2) + \sum_{\eta \in x_j} (R_\eta - \mu_j)^2/(2\sigma_j^2)\Big\}\Big]\Big) \\
&\times \exp\Big(\sum_{i=1}^{5} \Big[\theta_i \sum_{\xi \in y_i} s(\xi|x)^\mathsf{T} - \int_W \exp\left(\theta_i s(\xi|x)^\mathsf{T}\right) \mathrm{d}\xi\Big]\Big), \qquad (6.6) \\
&\theta_{ij} \in \mathbb{R}, R_\eta \geq 0.
\end{aligned}$$

The specification of priors for the θ_is is difficult. In regression models with a design matrix of full rank, a common choice are flat improper priors. In our situation, however, with improper priors on the θ_is we cannot guarantee a proper posterior since the statistics $t_j(y_i)$ have positive posterior probability of being zero (Waagepetersen 2005). It is also difficult to specify informative priors on the θ_is since we only have a qualitative understanding of these parameters. We choose to try out the values 2, 4, and 8 for the prior standard deviation σ and check to which extent the posterior results are affected by the choice of prior. It turns out that essentially the same posterior results are obtained with different values of σ; in the following we restrict attention to the results for $\sigma = 8$.

Monte Carlo estimates of posterior distributions are calculated using simulations from (6.6). We use a hybrid Markov Chain Monte Carlo algorithm (see e.g. Robert and Casella, 1999), where $\theta_1, \ldots, \theta_5$ are updated in turn, using random walk Metropolis updates, followed by a random walk Metropolis update of $\mathbf{R}$. The proposal distributions for these random walk

updates are multivariate normal with diagonal covariance matrices, i.e. matrices with non-zero entries in the main diagonal only. The vector of proposal standard deviations for θ_i is given by $k\hat{\sigma}_{i|\mathbf{y}}$, where k is a user specified parameter and $\hat{\sigma}_{i|\mathbf{y}}$ is an estimate of the vector of posterior standard deviations for θ_i obtained from a pilot run. The value of k was chosen to give acceptance rates for the MCMC algorithm around 25 %. The vector of proposal standard deviations for $\mathbf{R}$ is given by the vector of prior standard deviations divided by 2.

6.3.2 Results

6.3.2.1 Interaction parameters

The Bayesian approach does not yield fixed estimated parameters but instead a distribution of these parameters. We therefore cannot present these in a table similar to Table 6.3. However, we can calculate the probability $P(\theta_{ij} > 0|y)$, i.e. the probability that a particular parameter is larger than zero given the data.

Similar to Table 6.4, Table 6.6 lists the significant positive and negative interactions. A comparison of Table 6.4 and Table 6.6 shows that overall a larger number of parameters is significant in the Bayesian approach. Again, we observe many more negative than positive interactions; positive interactions only occur with seeder species 2.

Again, for a better general overview the lower plot in Figure 6.9 is a grey scale plot of the posterior probabilities $P(\theta_{ij} > 0|y)$ where dark grey scales are associated with large values of these posterior probabilities. The starred parameters are those for which 0 is outside the 95% posterior interval

Table 6.6: List of all significant positive and significant negative interactions for all resprouters for Bayesian approach using varying interaction radii

	positive interaction	negative interaction
1. *Alexgeorgia nitens*:	none	with all seeders
2. *Conostylis canescens*:	none	seeder 1, 4 and 5
3. *Dasypogon bromeliifolius*:	none	seeder 1
4. *Eremae astrocarpa*:	none	seeder 1 and 4
5. *Hibbertia crassifolia*:	none	seeders 1 and 3
6. *Hibbertia hypericoides*:	none	seeder 5
7. *Hibbertia subvaginata*:	none	seeder 5
8. *Hypocalymma xant.*:	none	seeders 1 and 2
9. *Lomandra sp.*:	none	seeders 1, 3, 4 and 5
10. *Lyginia barbata*:	seeder 2	seeder 1 and 4
11. *Phlebocaria filifolia*:	none	seeder 1 and 2
12. *Chordifex sinuosus*:	none	none
13. *Scholtzia involucrata*:	none	seeder 2, 4 and 5
14. *Allocasuarina humilis*:	none	seeder 3 and 5
15. *Banksia attenuata*:	seeder 2	seeder 5
16. *Banksia grandis*:	none	none
17. *Banksia ilicifolia*:	none	none
18. *Banksia menziesii*:	seeder 2	none
19. *Eucalyptus todtiana*:	none	seeder 1

of θ_{ij} given by the 2.5% and 97.5 % quantiles. In comparison with the upper plot in Figure 6.9, it is striking that the Bayesian approach seems to yield more clear-cut results than the maximum likelihood inference, since the intermediate grey scales are less frequent in the lower plot and there are more significant results. It is not directly obvious why this is the case. One might argue that a model with flexible interaction radii fits the data better. However, significance was determined differently in the Bayesian and in the frequentist approach. For the latter it is known (see Section 3.5.3) that the estimated standard errors are only approximate. Only a detailed simulation study could determine whether an appropriate type one control has been achieved in either of the cases.

Similar to the maximum likelihood results, resprouter 1 seems to have a clear repulsive effect on seeders, and this also seems to be the case for resprouters 2, 4, 5, 8, 9, and 14 (again we should exercise caution with resprouters 2, 9, and 14 where the prior may be concentrated on too small interaction radii). Resprouters 15 and 18 seem to have a distinct attractive effect on seeders. Looking at rows in the lower plot in Figure 6.9, seeders 1 and 4, for example, seem to be repulsed by resprouters 1-5 and 8-11 and attracted by resprouters 12, 15, and 18. The Bayesian analysis shows that it may not be valid to interpret all the θ_{ij}s with NAs in Figure 6.9 as corresponding to hard cores, since a number of these θ_{ij}s do not have strong posterior evidence of being different from zero.

In Section 6.4 these results will be discussed from an ecological perspective.

6.3.2.2 Interaction radii

The individual interaction radii are not of particular interest, so for resprouters $j = 1, \ldots, 19$ we just consider the posterior distributions of the empirical mean $\bar{R}_j$ and the empirical standard deviation s_j for the R_η with $\eta \in \mathbf{x}_j$. Table 6.7 shows prior means and standard deviations for the interaction radii and posterior means of the empirical means and standard deviations for the interaction radii associated with each resprouter type. Except for the very sparse resprouters 16 and 19, there is very little difference between the prior mean or standard deviation and the posterior mean of the empirical mean or standard deviation.

6.3.3 Model validation

In analogy with Section 6.2.5, denote by $L(r; Y_i, \theta, \mathbf{R})$ the estimate of the L function obtained from the point process Y_i using the intensity function corresponding to the interaction parameter vector θ and interaction radii $\mathbf{R}$. Following the idea of posterior predictive model checking (Gelman et al. 1996) we consider the posterior predictive distribution of the differences $\Delta_i(r) = L(r; y, \theta_i, \mathbf{R}) - L(r; Y_i, \theta_i, \mathbf{R})$, $r > 0$, i.e. the distribution obtained when $(Y_i, \theta_i, \mathbf{R})$ are generated under the posterior predictive distribution given the data y. If zero is an extreme value in the posterior predictive distribution of $\Delta_i(r)$ for a range of values, we may question the fit of our model. In practice, we generate a posterior sample $(\theta_{i,1}, \mathbf{R}_1), \ldots, (\theta_{i,m}, \mathbf{R}_m)$ and for each $(\theta_{i,k}, \mathbf{R}_k)$ we generate new data $y_{i,k}$ from the conditional distribution of Y_i given $(\theta_{i,k}, \mathbf{R}_k)$. We can then estimate the posterior predictive distribu-

Table 6.7: Prior and posterior means and standard deviations for the interaction radii and posterior means across the empirical means and standard deviations of the distributions of the interaction radii associated with each resprouter species

	prior mean	posterior mean of $\bar{R}j$	prior std	posterior mean of s_j
1. *Alexgeorgia nitens*:	25	25	7.5	7.5
2. *Conostylis canescens*:	10	10	2.5	2.5
3. *Dasypogon bromeliifolius*:	37.5	37.2	11.25	11.5
4. *Eremae astrocarpa*:	47.5	47.5	11.25	11.3
5. *Hibbertia crassifolia*:	17.5	17.4	3.75	3.7
6. *Hibbertia hypericoides*:	15	15.0	2.5	2.5
7. *Hibbertia subvaginata*:	17.5	17.4	3.75	3.7
8. *Hypocalymma xant.*:	17.5	17.4	3.75	3.8
9. *Lomandra sp.*:	6.6	6.6	20	20
10. *Lyginia barbata*:	60.0	60.0	20	19.9
11. *Phlebocaria filifolia*:	25	25.0	2.5	2.5
12. *Restio sinuousus*:	50	50.0	12.5	12.5
13. *Scholtzia involucrata*:	40.0	40.0	5	5
14. *Allocasuarina humilis*:	90.0	90.2	20.0	13.3
15. *Banksia attenuata*:	275	273.8	62.5	71.7
16. *Banksia grandis*:	125	59.7	37.5	0.00
17. *Banksia ilicifolia*:	37.5	39.5	37.5	26.8
18. *Banksia menziesii*:	150	147.6	50	51.6
19. *Eucalyptus todtiana*:	130	172.5	60	0.00

tion from the sample $L(t; y_i, \theta_{i,k}, \mathbf{R}_k) - L(t; y_{i,k}, \theta_{i,k}, \mathbf{R}_k)$, $k = 1, \ldots, m$; we used $m = 100$ approximately independent simulations by subsampling from a Markov chain of length 200,000.

Figure 6.11 presents estimated upper and lower boundaries of the 95 % posterior intervals for the posterior predictive distributions of $\Delta_i(r)$, $r > 0$, for the 5 seeder species.

These intervals take into account the uncertainty of the model parameters θ and $\mathbf{R}$ and are quite wide in comparison with the envelopes in Figure 6.10. The wide envelopes probably arise because of the posterior uncertainty regarding the interaction radii; the intensity function at a seeder location may a posteriori be very variable if it is highly uncertain whether the seeder location falls within a resprouter influence zone or not.

Nevertheless, the width of the envelops seems to be contradictory to the fact that the results for the Bayesian approach are more clear-cut than those for the frequentist approach (see Figure 6.9). To ensure that these wide envelopes are not resulting from an error in the program, they were independently recalculated twice resulting in the same wide envelopes.

In accordance with the results in Section 6.2.5, there is evidence of clustering for seeder 1 and indication of repulsion at small distances less than 20 cm for seeder 4. In contrast with Section 6.2.5, the posterior predictive intervals for seeder 2 indicate clustering. As for seeder 1, this may be explained by offspring clustering around locations of parent plants. Another explanation is inhomogeneity not accounted for by our model, since seeder 2 plants are absent in the top part of the observation plot, cf. Figure 6.8. Also, for seeder 3 there is evidence of clustering. The posterior predictive intervals

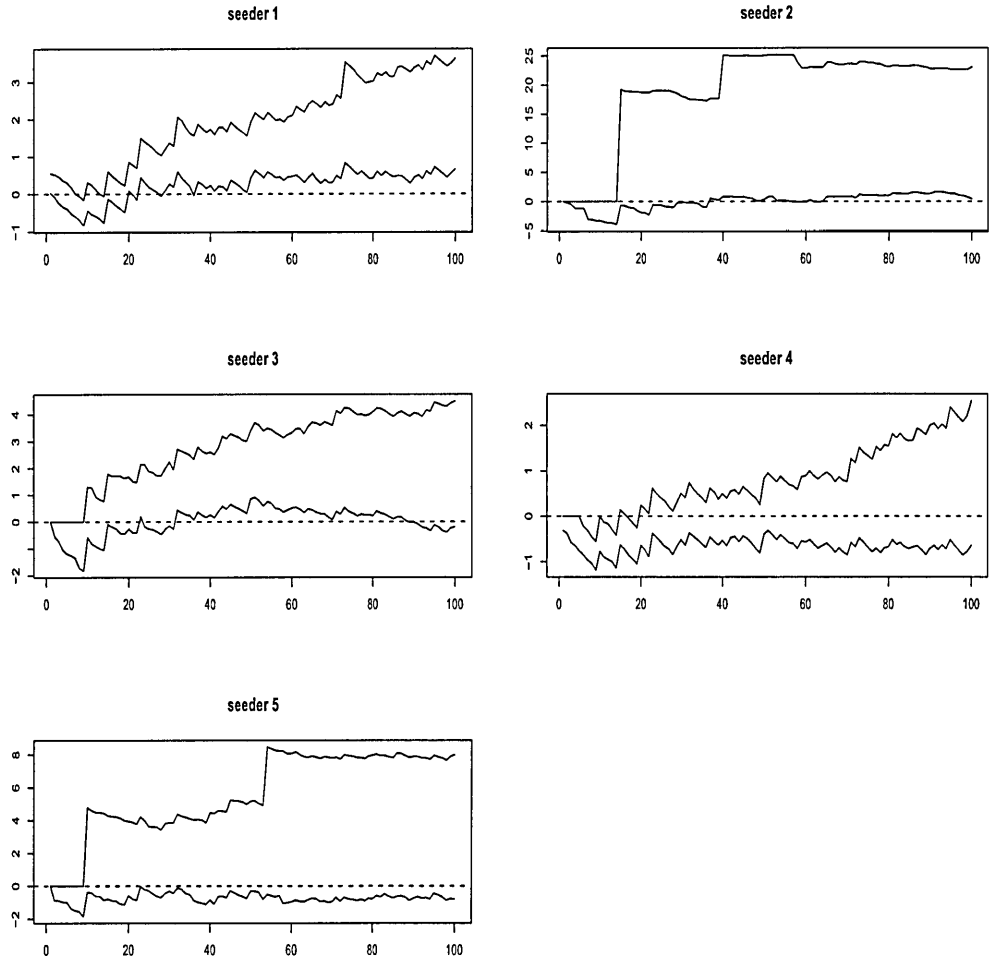

Figure 6.11: Upper and lower boundaries (solid lines) of the 95% credibility interval for the posterior predictive distribution of $L(r; y, \theta_i, \mathbf{R}) - L(r; Y_i, \theta_i, \mathbf{R})$ for seeders 1-5. Distance r is in cm.

for seeder 5 do not provide evidence against our model.

In order to verify the convergence of the MCMC algorithm, we produced trace plots of the parameter distributions for the interaction parameters. These indicate that the algorithm has converged in the 20,000 iterations.

Refer to Appendix C for these plots. We do not display all convergence plots for the initeraction radii as this would involve another 3168 plots.

6.4 Discussion

From a statistical point of view, our analysis shows the difficulty of modelling spatial interactions in a plant community which requires very complex models with a large number of parameters. In this situation, we found the Bayesian approach more useful than the frequentist approach as it allowed a more flexible and realistic model. Furthermore, in our analysis taking biological background information into account yielded a hierarchical model. However, we are aware that the model does not sufficiently capture all interactions that may be present in the dataset. It does not consider intra-species interactions for the seeders and, similarly, assumes that the seeder species are independent given the resprouters. Incorporating all these aspects into a single model, though, is likely to be computationally intractable.

From an ecological perspective, we were able both to confirm existing knowledge on species' interactions and to generate new questions and hypotheses for applied researchers on species' interactions that were previously unknown.

As in Chapter 4 and Chapter 5 we will discuss the methodology, the implication on the theory of biodiversity as well as the results for the current data set in turn.

6.4.1 Discussion of methodology

In the context of ecological plant communities, most notably when considering one with a high degree of biodiversity, one should aim at fitting a model that is as simple as possible and where parameter estimation is straightforward. An obvious choice of model would have been the Continuum Ising model, in particular since we showed that existing standard software may be applied for an approximate parameter estimation. However, the model turned out to be too simple as it ignored that symmetric interaction is highly unlikely in any ecological context. This is even more problematic here, where species with two very different fire regeneration strategies are being modelled and mutual interaction is very unlikely. Nevertheless, the estimation method developed for the Continuum Ising model may be of use in other contexts where mutual interaction can reasonably be assumed and thus has been included in this thesis.

As a consequence of the inappropriateness of the Continuum Ising model for the given context, we developed a model which takes the information on regeneration strategies into account. We did not consider interactions that are likely to be very close to zero. Thus, even though the number of potential interaction parameters would have been increased when asymmetry is included in the model we could use background information to reduce the number of parameters that had to be estimated from the data.

We consequently devised another multivariate model which assumes that the seeders are neither encouraging nor inhibiting the growth of the resprouters whereas the resprouters can interact positively or negatively with the see-

ders. Through assuming a fixed interaction radius for all individuals of a given resprouter species we applied a similar estimation method as for the previous model this time using the likelihood rather than the pseudolikelihood. However, the fit of the model was not always acceptable, which might be due to the assumption of a fixed interaction radius.

In a next step, we thus extended the model allowing for the interaction radius to vary among the individuals within the five species. We used a Bayesian approach here and apply MCMC methods for parameter estimation. This much more flexible setting allowed the estimation of the distribution of 3456 parameters. As noted, the Bayesian approach seems to yield more clear-cut results than the maximum likelihood inference, so varying the interaction radii has apparently made an impact on the quality of the estimation. However, the estimation process involved for the Bayesian approach turned out to be more tedious. It was not possible to use standard software for parameter estimation and the program took a very long time to run – a run of 200,000 iterations took 2 weeks, whereas previous estimation algorithms for the simpler models had taken only a few minutes to finish.

We currently assume that there are no resprouter-resprouter interactions and no seeder-seeder interactions. This clearly ignores intra-specific interaction, which we know exists for some of the species, see Section 4.4.4. A revised model should include these interactions.

Further, the application of the current approach is not restricted to the dataset described here or to datasets collected in environments with two different fire- strategies. For instance, a similar modelling approach could be envisaged when the spatial pattern of annual and perennial plants in

moderate climates is modelled, with the perennials taking on the role of the resprouters and the annuals that of the seeders. This would involve the modelling of the locations of the annuals conditional on the locations of the perennials.

6.4.2 Discussion of biodiversity theory in this context

Ecologists are particularly interested in revealing how individuals interact and whether this interaction varies between species since previous results have shown that a mixture of positive and negative interactions might promote biodiversity (Hopper 1979). Our results clearly indicate that resprouter-seeder interaction may be both negative and positive and that the same species can have both positive and negative inter-species interactions. So clearly the identity of the neighbour does matter which may be an indication in favour of niche theory (Uriate et al. 2004).

6.4.3 Discussion of results for data set

The spatial pattern formed by plant species in the heathlands of Western Australia has not been analysed and modelled in any previous study. The only study that has taken the spatial pattern formed by individuals in the area is the work by Enright et al. (2004), which we have mentioned earlier in Chapter 4 but this paper only applied exploratory spatial point process methods and does not fit a model to the data. As a result we cannot compare the results obtained here to other results in the literature. The interpretations given below are the result of discussions with experts and are of a very speculative nature (Dixon 2005; Armstrong 2005).

Some of the interactions revealed here could be explained from ecological background knowledge. For instance, the negative interaction of the most abundant resprouter *Alexgeogia nitens* might be a result of the dense root mat formed by this species making it very difficult for plants from other species to establish themselves close to them. On the whole, *Alexgeorgea nitens* clearly is a very dominant species. It has the highest abundance in the dataset and is the only resprouter species that has negative interactions with all seeders.

Similarly, the strong positive interactions between the resprouters *Banksia attenuata* or *Banksia menziesii* and seeder *Astroloma xerophyllum* may be explained by specific associations between soil fungi and the plant roots of the seeder. These associations, termed ericoid mycorrhiza, facilitate nutrient uptake from the soil by plants from the *Ericaceae* family. They enable the seeder to use very complex organic material, which is normally impossible to extract from the soil. *Banksias* produce a very dense and expansive sub-soil ramification of cluster roots or proteoid roots, see Section 2.4.6. These cluster roots survive only one growing season but are highly resistant to microbial decay indicating that they are composed of materials difficult for regular microbes to break down. However for the specialised *ericoid* mycorrhiza this may present fewer difficulties due to the specialised enzyme systems associated with these fungi that enable access to complex organic assemblages (Cairney and Ashford 2002). As a result, individuals from seeder 2 find more nutrients close to resprouters 15 and 18, leading to a positive interaction. As all seeder species apart from species 3 (*Conospermum crassinervium*) in the analysis are from the *Ericaceae* family, other positive interactions between

these seeder species and species from the *Banksia* family (resprouters 15 to 18) might be explained in a similar way.

However, for a large number of interactions no biological explanation could be found. There clearly is a need for further biological research to understand these aspects and to yield a better understanding of the overall community dynamics.

From a more general point of view, knowledge about the direction and strength of interaction will aid re-naturation of disturbed areas. Ecologists will now be aware of which species thrive in each other's vicinity and which do not. When an area is replanted this may be taken into account, irrespective of whether the reason for the interaction is known. In an application, a subarea of the area due to be mined could be modelled using an approach similar to the model described here in order to provide information for the larger area, assuming that data collection for the entire area might be infeasible.

Chapter 7

Discussion

Here, we summarise and critically discuss the results obtained with the methods applied and developed in this thesis. In Section 7.1 we discuss the methods and their suitability in the context of plant community ecology as well as the ecological insights gained from the approaches, both for the specific data set and, more importantly, for research into biodiversity in general. Section 7.2 addresses in detail scope for further work.

7.1 Conclusions

This thesis investigated whether spatial point process modelling techniques can be suitably applied to ecological communities to determine factors that promote biodiversity. To this end, three approaches were taken to overcome the main limitations in current methodology outlined in Section 2.3.3:

- Parsimonious explorative statistics suitable for a complex point process data set were developed in Chapter 4.

- The dimensionality of the data set was reduced by concentrating on the main aspects of spatial behaviour in Chapter 5.

- A parsimonious spatial point process model was fitted to the multi-species spatial point pattern formed by individuals in Chapter 6.

We discuss a number of technical aspects related to the methodology and the contribution to statistical knowledge made in this thesis (Section 7.1.1), summarise its contributions to the analysis of plant community dynamics and the understanding of biodiversity in particular (Section 7.1.2), investigate the quality of the contribution spatial point process methods as statistical methods have made (Section 7.1.3) and close with a general outlook on spatial modelling in ecology (Section 7.1.4).

7.1.1 Technical aspects of the methodology

Previous applications of spatial point process analyses have been restricted to very small numbers of points and small numbers of types of points, at most three species in an ecological context. Furthermore, most research within the context of spatial point processes has been theory-driven (Diggle 2003). This implies that potentially useful methods are often developed initially without an existing data set or a research question related to this data set in mind. This is done to provide a mathematically tractable model for potential point pattern data. Consequently, these methods have been applied to an available data set purely to demonstrate that parameter estimation was feasible and that data existed to which the model may be fitted. Typically therefore, models were neither constructed to answer explicit applied questions nor

has the models' relevance to actual problems been assessed. This has had clear consequences for the methodologies available to the applied researcher leading to a number of shortcomings that so far have not been looked at in detail. These may be summarised into two main issues:

a) lack of suitability of methods to multivariate setting

b) assumptions of theoretical models that simplify mathematical description but are unrealistic in many real-life studies.

These problems are reflected in all aspects of the existing methodology and we outline how this thesis has addressed these for each of the three approaches described in Chapters 4, 5 and 6.

Exploratory data analysis When analysing only a small number of species, manual inspection of exploratory data analysis plots is not particularly time-consuming. This becomes intensive when highly multi-species data sets like the Cooljarloo data set (Armstrong 1991) are analysed or even more so in the case of the rainforest data sets from the CTFS network (Burslem et al. 2001). The first order summary statistics, where the decision as to whether a data set is sufficiently homogeneous to be modelled by a particular model, has often been done manually (refer to a) above), which is time-consuming as well as subjective. A formal statistical test could be automated and would provide an objective decision criterion. However, a simulation study suggested that the sole test available in the literature was less powerful than two approaches developed in this thesis. In particular, when no prior knowledge of potential areas of high plant density is available the test proved

to lack practical relevance. This thesis has explored methods that can be applied easily to data sets to test homogeneity assumptions formally and efficiently (refer to Section 7.2 for a more detailed discussion of this issue).

In addition, first-order summary statistics, have typically applied simple kernel estimation procedures to date. Here, the crucial decision on the size of the bandwidth has been rather subjective. The exactness of intensity estimation and the resulting values have previously not been of particular interest, as the estimated surface was merely used to provide qualitative support for the assumption of homogeneity such that a model that assumes homogeneity could be fitted (and thus concerning b) above). Only recently have the estimated values been explicitly used when inhomogeneous second order summary statistics have been applied on the basis of the estimated intensities using simple kernel estimation methods. More sophisticated intensity estimation methods that use different types of bandwidths depending on the sparseness of the data have not been applied hitherto. Furthermore, the influence of the quality of the first order estimation on second order estimation has been assessed.

Similarly, with reference to a) above, the use of second order summary statistics together with relevant simulation envelopes is straightforward for the analysis of a small number of subpatterns. In a multivariate setting however, the process of deciding whether a pattern exhibits complete spatial randomness or not can become tedious. Hence, a parsimonious testing procedure based on CUSUM approaches was developed in this thesis based on second order summary statistics, in particular L-functions. In addition to the approach being quick to apply, a simulation study suggested that the

CUSUM method is more powerful than standard methods when testing the null hypothesis of complete spatial randomness against clustered and regular alternatives. The approach has the potential of being extended to more general settings. It could be applied to pairs (or larger groups) of species, i.e. to the paired L-functions or other second order summary statistics such as the pair-correlation function (refer to Section 7.2 for a more detailed discussion of this issue).

Principal component analysis Previously, due to a lack of highly multivariate data sets, there was little need to summarise data sets with large numbers of species. Similarly, there was little interest in identifying those aspects of spatial behaviour that the multiple subpatterns varied in most and or in characterising the spatial behaviour of the combined pattern. As a result, multivariate statistical methods for spatial point pattern data have not be developed hitherto (refer to a) above). This thesis developed a principal component analysis based on functional principal component analysis (FPCA) that, together with a standard hierarchical clustering analysis yielded groups of species with similar spatial behaviour and thus characterised the spatial behaviour of the community. A feasibility study assessed the capability of the methodology to distinguish between patterns of known spatial characteristics and provided advice on the choice of the second order summary statistic. A simulation study demonstrated the methodology's robustness in the presence of errors common in ecological applications.

It has to be acknowledged, though, that there is a strong dependence of the quality of the grouping on the quality of the first and second order

estimators, and on the consequent smoothing of the estimated second order summary statistics. As a result, more accurate first order estimators may improve multivariate methods. Similarly, only simple smoothing methods (B-splines) have been used. More sophisticated smoothing methods may provide even better results and, more specifically, be able to cope with the apparent estimation bias in inhomogeneous L-functions at larger distances (see Section 7.2).

Modelling approach So far, explicit models have not been applied to data sets with a large number of species (refer to a) above). As a result, there has been no need for parsimonious models and parameter estimation. Similarly, although particular models such as the Continuum Ising model may be generalised to a multivariate setting, this has only been done theoretically without an application to data (refer to b) above).

The complexity of the modelling task and hence the inherent complexity of any potential model leads to complexity at all stages of analysis. For instance, plotting all subpatterns in one plot is feasible when only two or three species are analysed by using different colours and reveals valuable information on inter-species interaction. However, it is difficult to do something similar if 67 species or even several hundred subpatterns as in the CTFS rainforest data sets are analysed (recall Figure 2.6). As a result, even standard methods of graphical data representation used in simpler problems have to be reconsidered.

Likewise, parameter estimation becomes more complex. In this thesis, we have devised a parsimonious procedure that may be used for parameter

estimation in a multi-type Continuum Ising model. Despite that fact that the model was too simple for the study data set such that we moved on to a model considered more appropriate, the estimation method itself is of interest and may be applied in other situations. A similar estimation approach was used when a hierarchical model in a frequentist setting was applied that was deemed more suitable for the data set but made the assumption of fixed interaction radii for the individuals within each species. Though this assumption may well be realistic in other contexts, it was considered too restrictive for the current application.

Hence, the hierarchical model was generalised in a Bayesian setting to allow for intra-species interaction radii variation. Here MCMC methods were applied for parameter estimation where the inherent complexity of the data set led to long running times. The MCMC estimation of the parameters in the model considered here took about two weeks to run, which is not prohibitive. However, MCMC estimation requires a substantial amount of fine-tuning with regard to both prior distributions and proposal distributions to improve convergence. With complex models and the resulting long running times, this part of the estimation process may take a very long time before parameter estimation may begin. Nevertheless, the Bayesian approach turned out to be very flexible, particularly with regard to the estimation of the interaction radii. In a non-Bayesian context these parameters could not have been feasibly estimated.

The modelling approach taken here, in particular in the Bayesian setting, is highly appropriate for future applications. For instance the analysis of the rainforest data sets mentioned in Section 2.3.2 will benefit from the flexibility

of the parameter estimation, as it allows the fitting of appropriately realistic models despite the models' complexity (Best et al. 1996). The rainforest data sets contain mark information on the trees and information on soil variables. This allows analysis of habitat associations of species but results in more complex modelling requiring suitable estimation methods. In Section 7.2, we discuss this in further detail.

Further, the interpretation of the results also becomes extremely complex in the same way as the mere display of the results becomes difficult. For example, in many papers where MCMC methods have been used convergence plots for each parameter are presented to indicate that the MCMC algorithm was run long enough to yield reliable estimates. For the model here this would have involved 100 convergence plots for the interaction parameters and 3168 plots for the interaction radii. In Section C in the appendix we have thus only presented the plots for the interaction parameters.

To date, there has been no work investigating how information from the FPCA methodology may be used to inform modelling. It is difficult to say how this should be done. Initially, one might suggest not distinguishing between species that have been categorised into the same group and treating them as one "meta-species". However, we do not feel the grouping justifies this, since such a procedure might obscure existing patterning. For instance, the joint pattern of two or more clustered patterns might result in a pattern that does not exhibit any clustering. As a result, the resulting "meta-species" may exhibit a different type of spatial patterning than the original two or more patterns.

We thus propose that species in the same group should instead be modelled with a similar spatial point process model. This is of particular interest, since the grouping has been done on the basis of the largest variation in spatial patterning among the species. Through this, the most important spatial components are identified. If this grouping is used to inform model choice, the chosen model is then more likely to explain large parts of the overall variation in the data set. This may lead to the identification of those underlying processes that have the largest impact on community dynamics.

In Section 7.2, we will indicate further work to achieve this and to improve the methodology in general.

7.1.2 Application of spatial point process modelling to plant communities and contribution to biodiversity theory

The previous sections summarised the statistical advances for spatial point process modelling in the context of a large number of types of points, i.e. in this thesis large numbers of different species. The results obtained when the methods were applied to the Cooljarloo data set may also be interpreted to contribute to knowledge on biodiversity. In Section 2.3.1.2, we introduced two opposing theories, niche theory and neutral theory and in Section 2.3.2, we highlighted potential contributions of spatial point process modelling to linking these theories to observed biodiversity. We now summarise how our results may indicate which one of the theories is more likely to hold for this data set.

Interaction strength and neighbourhood identity Both the exploratory data analysis and the principal component analysis revealed that intra-specific interaction varies among species. From the exploratory data analysis and both the frequentist and the Bayesian approach to parameter estimation we learned that inter-species interaction varies as well. This indicates that interaction strength and direction varies with the identity of the species involved since negative, positive and no interactions were found. This argues in favour of niche theory, which claims that it makes a difference which other species a certain species interacts with.

Similarly, niche theory would predict low and neutral theory higher interaction strength. This is because neutral theory postulates that community dynamics are mainly based on interactions rather than on species' identities and properties. It is not possible to quantify what should be understood as high interaction strength from ecology theory. However, we may interpret a statistically significant interaction strength as high, as it apparently has had an effect on the spatial patterning of the community. We have found both significant and non-significant interactions. It is not entirely clear whether this is an indication that niche theory is valid for some species and neutral theory for others, or whether it suffices to find some significant interactions to provide evidence in favour of neutral theory.

Further, niche theory postulates varying interaction strength among different species whereas neutral theory would assume that the interaction strength of a particular species is the same no matter which other species is involved. We find that interaction strengths and ranges clearly vary among the species and the interaction ranges even among the individuals. Both interaction

strengths and ranges are also clearly linked to species' properties, such as their size and their identity such that we are unable to assume that it is the interaction among the plants rather than their identities and properties that rule community dynamics. This, again, provides evidence in favour of niche theory.

Microhabitat specialisation Knowledge about individual species' microhabitat associations would provide evidence in favour of niche theory as it predicts that species have adapted to specific growing conditions in order to coexist. However, on the basis of the given data set, we are unable to draw any conclusions on potential microhabitat associations of the involved species (see Section 2.3.2). This would require data on these microhabitats, e.g. on soil nutrients and shading (canopy density) such that these could be incorporated in a model.

Nevertheless, for the study data set we can assume that the growing conditions are homogeneous throughout the plot and interpret the resulting pattern as primarily resulting from intra- and inter-specific interactions. Using the principal component methodology developed in this thesis, groups of species with similar intra-specific behaviour could be identified. More importantly, we could establish those aspects of spatial behaviour that vary most among the species. This is interesting because we are dealing with a data set of a truly ancient community, which has established its dynamics over thousands of generations in an area which has undergone substantial climatic changes as a result of continental drift. Many of the species in the area have been around for millions of years (Dixon 2005) and we may thus

assume that we are dealing with a very stable community that has adapted to the given circumstances in an optimal way resulting in the species developing niche behaviour that is crucial for stable coexistence. It could thus be concluded that the spatial behaviour at small distances constitutes a niche-behaviour, which may support coexistence of the large number of species in the area under investigation.

Density dependent processes The exploratory data analysis revealed that a large proportion of the species exhibits clustering. The Janzen-Connell hypothesis predicts that species' survival is improved with increasing distance from the location of a conspecific adult tree and decreases with the number of conspecific individuals in close vicinity. This would predict that older individuals would tend to form a more regular pattern than juvenile individuals of the same species that might be clustered as a result of initial seed dispersal.

As indicated in Chapter 4 we cannot compare the spatial patterning of juvenile and old species due to the unique age structure in the particular community. Further, the study data set was collected on a 22m by 22m plot only. Processes predicted by the Janzen-Connell hypothesis are likely to operate on a much larger scale such that evidence against or in favour of the hypothesis cannot validly be found in the data set.

We did, however, reveal that some resprouters, after having survived for extremely long times, still exhibit spatial clustering. This may provide evidence against the Janzen-Connell hypothesis for the given system. However, spatial point process methods may contribute to the discussion in a more detailed way if data on the age of the individuals were available. More spe-

cifically, the PCA methods developed in this thesis may be applied to the pattern of the several species with each species split into different age groups, as available e.g. for the rainforest data. If different age-groups of the same species are allocated to groups with different spatial behaviour, and most notably with different degrees of clustering, this may support the Janzen-Connell hypothesis.

On the whole, this analysis has found more evidence in favour of the niche theory as opposed to neutral theory. So, apparently population dynamics are affected by the species' identities and hence a loss in biodiversity influences these dynamics. This ultimately indicates that for the system considered here loss in biodiversity may be linked to ecosystem functioning.

More specifically, our results contradict Chave (2004) who supposes that niche theory may rather be used to explain coexistence in species-poor communities and neutral theory applies to species-rich communities.

7.1.3 Wider benefits of a spatial statistical approach

We have applied spatial point process models, i.e. statistical models that model the location of objects, based on background knowledge about underlying mechanisms that may have caused the observed patterns, in the context of a biodiverse plant community. Through this, we were able to gain a better understanding of the particular community. More importantly, the approaches we have taken here could contribute to ecological theory as outlined above.

In this section, we summarise the direct benefits that we have gained from taking a statistical modelling approach as opposed to applying an individual

based model (IBM). In doing so, we refer back to Section 2.1.6 where we made a case for applying statistical modelling approaches in the context of plant community dynamics.

A statistical model provides an overall picture of the dynamics in the system by still taking each individual into account and by summarising the characteristics of the spatial structure in the community using a model based on parameters. Through the modelling approach we were able to capture the overall properties of the whole community in a small number of estimated parameters (frequentist approach) or their distributions (Bayesian approach). The number of parameters and hence the complexity does increase with the number of species involved but it does not increase with the numbers of individuals modelled which typically is the case in an IBM. This ultimately improves a model's interpretability.

In general, a statistical approach provides clear decision criteria. For instance, we were able to analyse the spatial pattern formed by individual species in the explorative data analysis and assess which of the species patterning differed significantly from a random pattern. Similarly, we summarised the spatial behaviour in the plant community across species, when we applied FPCA to the spatial point pattern and determined the most distinguishing aspect of spatial patterning. In the given data set, this turned out to be clustering at close distances. This implies that the most distinctive feature in spatial behaviour among species in this community is their behaviour at close distances and indicates a niche behaviour, as detailed above.

The principal component approach yielded a projection of the complex properties of the spatial pattern into a lower-dimensional space. In addition,

the approach allowed the choice of dimension of this space based on objective methods, i.e. the percentage of variance explained by the principal components, employed in this thesis. We could even identify which of the lower dimensions is the most influential since principal component analysis returns the principal components in order of importance.

Since spatial point process models are statistical models they can be used to predict the resulting patterns given the model parameters. Statistical models may be used for inference since an error term, i.e. a random component, is typically included in the model. This random component accounts for variation in the data that cannot be explained by the processes in the system that are assumed to be important for community dynamics and thus to have impacted on the spatial pattern. It is thus possible to quantify the relative importance of the random and the systematic components. In particular, we could determine which parameters, and hence which resprouter-reseeder interactions, are statistically significant, i.e. have made a meaningful contribution to the observed pattern. Through this, we may conclude that these are the interactions that are important in the community dynamics of the system.

As a consequence of using a statistical approach, we also have a mechanism that allows the assessment of the suitability of our models. We may assess the fit of a model to the data, both in the study data set but more importantly in other, more complex data sets. In the special case of the final Bayesian model we are able to say that the model fits some of the seeders' patterns rather well (in particular seeder species 5, *Leucopogon striatus*) and others less well. This clearly suggests that modelling has to be improved

for some of the species. So, on a more general level, by indicating that a given model is not yet suitable enough for a complete understanding of the system, statistical approaches thus encourage new research; any model is not considered as the final or ultimate model but a model in a transient phase whose weaknesses have clearly been identified.

Due to the lack of data on further properties of the plants, such as their height, we could not infer interaction ranges of the individuals from the data. A Bayesian approach was applied, which did not assume equal interaction radii across the individuals of a particular species. This resolved the problem that we could not explicitly model an individual's interaction radius on the size of this individual. This may be seen as a clear benefit of the Bayesian approach. However, in addition this approach has equipped us with a means of taking inter-individual variation into account even in the presence of information on species' properties, which to date has only rarely been accounted for in plant community modelling. Thus, even if data on relevant plant properties were available a more refined Bayesian approach may be applied to account for inter-individual diversity.

Overall, this thesis has provided a number of statistical tools that may also be successfully applied in other situations where highly multi-variate data sets of spatial patterns occur. The methods are suitable in the context of plant community dynamics and have made contributions to the development of existing ecological theories on biodiversity. More specifically, they have also informed on the intra- and inter-species interactions that have an influence on the community dynamics in the study data set. For instance, the assumptions of positive interactions between species based on mycor-

rhizal associations could be formally confirmed. Further, knowledge about the interaction structure will help the rehabilitation of the area and the methods could be applied to provide similar information for other areas and thus contributing to conservation.

Spatial point process methodology and in particular the methods developed in this thesis may be applied to many other data sets. This includes other biodiverse plant communities or disturbed communities following a fire or other disturbances. More generally, they may be successfully used in the context of animal populations modelling the locations of e.g. bird's nests, fox lairs and similar data. Applications outside biology also exist in medicine, e.g. when the occurrence of certain types of cancer are analysed or in geology, e.g. when maps of different minerals exist. Through the simulation study in Chapter 5 we have also provided a means to inform applied researchers what level of exactness is required to make best use of the principal component analysis. In practice, this information may be valuable as it determines the necessary sampling effort.

All methods developed here, clearly have the potential to be generalised to be usefully applied in more complex situations. A working group has been formed which has started extending the approaches developed in this thesis to more complex data set, in particular the rainforest data sets from the CTFS network mentioned in Section 2.3.2.

7.1.4 Outlook for further modelling of space in ecology

7.1.4.1 Considering spatial structure

In Chapter 2 we emphasised the importance of considering spatial structure in the context of ecological modelling and, more specifically, in the context of plant communities due to the plants' inability to move and, hence, their reliance on local growing conditions. In particular, we pointed out that it is necessary to construct plant community models that are based on the individuals in a community rather than on their average abundance.

Using the approaches taken in this thesis, we were able to provide further evidence and thus strengthen the argument. We confirmed for example that for this data set the pattern formed by a large proportion of the species in the study data set is clustered rather than random or regular. As a consequence, being typically surrounded by other conspecific individuals will render the environment of the individuals clearly different from the average environment.

Furthermore, we found evidence that individuals interact, with varying interaction radii among species and individuals. Outside the interaction radius only very little interaction takes place, underlining that these interactions are local and determined by the respective local growing conditions. On the whole, we could provide further evidence that the mean field approach is very likely to be invalid for this plant community. This important observation is certainly not new but provides further justification for considering individuals in a spatially explicit approach and thus applying spatial point process methods and individual based models.

7.1.4.2 Spatial point processes or IBMs?

In Section 7.1.3 we have detailed the benefits of a statistical modelling approach and contrasted it against IBMs. Which of the two approaches, however, may be considered more appropriate in the context of plant communities spatial point processes or IBMs? We have detailed the general benefits of spatial point process models in 7.1.3 and concluded that the methods provide clear decision criteria and are suitable in the context of plant community dynamics and biodiversity. Nevertheless, spatial point process models only provide insight into ecological phenomena about which data exist. The current data set was collected at one point in time only, completely ignoring temporal processes. Whereas it is definitely possible to extend the approach to include a time-dimension (Diggle 2003) it requires, however, that data have been collected at several time points, which is very time consuming or might be infeasible. In the case of the type of plant community from which the Cooljarloo data set was collected, processes take place at a very slow pace. More generally, some developments within communities would require data collection over a very long time span such as several generations, which is highly impractical and might be impossible. If one were, for instance, interested in observing the development of the community over several generations or even in understanding phenomena such as speciation, etc., it may be impossible to collect data for this. However, using IBMs it is much more straightforward to simulate data for very long time spans and observe the development of a plant community.

On the other hand and perhaps more importantly, data collected at several points in time can still only yield results at the time-scale the data have been collected. Processes that may be important and may take place at a very fine temporal scale with autocorrelations may not be captured by the data due to practical data collection limitations. IBMs, however, assume continuous processes which are more likely to occur in nature than the rather discrete time approaches that the point process approach can yield.

As a result of the above arguments, we suggest, that it might be very useful to combine the deterministic or mechanistic approach taken by the IBMs and the spatial statistical approaches, in particular spatial point process models. IBMs can capture more details of the underlying processes and spatial point processes models processes yield an overall description characterising the pattern. Ideally a combination of the two approaches would yield a combined model of both the processes and the pattern. This might be done by applying state space models which have been successfully applied to model the dynamics of wild animal populations (Buckland et al. 2004) but need to be expanded as to-date they are not using an individual-based approach.

In a recent paper, Schneider et al. (2006) apply a spatial statistical approach using Bayesian estimation methods for interaction parameters in a mechanistic growth model, which provided insightful results on the shape of the interaction function for a specific data set. We are not aware of any other attempts of combining the statistical approach with a spatially explicit mechanistic model, in particular not of an application of spatial point process models.

Furthermore, IBMs would usually be fitted by assuming relevant mechanisms and properties of the plants such as size, etc. are then estimated from data. However, little is known about properties of community dynamics in general, such as interaction strength and interaction ranges (Bolker et al. 2003; Schneider et al. 2006). What is more, these cannot be directly observed in a specific data set. Hence, parameters representing interaction strength and direction derived from real data cannot be included in an IBM. Due to the inferential nature of the statistical approach, spatial point process models could be used to estimate these parameters, and consequently form part of an IBM.

In addition, the statistical models may be used to determine the relative importance of suspected mechanisms such as intra-and inter-specific interaction of the various properties of individuals. As a result, it would be possible to concentrate on only the relevant mechanisms or weigh mechanisms by importance in an IBM. As an example, an IBM could be constructed from the parameters estimated in this thesis in order to observe how the pattern might evolve over time.

Further, we have suggested in Section 2.5 that taking species' spatial behaviour before disturbance into account may improve rehabilitation processes. A study which compares rehabilitation success with and without considering the spatial behaviour may support further this argument and also inform about the size of the impact of considering spatial behaviour. However, such a study would be costly and would have to be run over very many years due to processes taking place at a very large timescale. IBMs may be a very useful tool to simulate communities after mining or other

disturbances.

7.2 Scope for further work

Apart from combining the spatial statistical approach and the IBM approach which may provide ample scope for further work for modellers, spatial point process methodology itself may be improved in many aspects. This thesis is only the first step in the development of spatial point process approaches suitable in the context of plant communities and much work remains. Here, we indicate how the methodology could be improved and generalised to more complex data sets.

7.2.1 Exploratory data analysis

The quality of first and second order summary statistics estimation is crucial for the entire modelling process, as these, and most notably the intensity estimation, have an influence on all further steps. Summary statistics are applied repeatedly throughout the process, even up to the final model validation process, as described in Chapter 6. However, to date intensity estimation has mainly used basic kernel estimation methods. These employ the same bandwidth to the whole pattern regardless of local point densities. As a result, areas of high intensity may be over-smoothed and local structures may be missed out. This is clearly a disadvantage as it is the local structures in which we are most interested. Here, it might be advisable to apply local kernel estimation approaches with a varying bandwidth, such as adaptive kernel estimation (Brewer 2000). In addition, spline smoothing approaches might be more suitable. Furthermore, a model based approach to intensity estimation, e.g. using regression spline mixed models (Rice and Wu 2001),

might be another useful approach as the error term in the model might yield a decision criterion as to the choice of bandwidth.

Clearly, applying the Kolmogorov Smirnov test in order to assess the inhomogeneity of a spatial pattern is only one possibility of improving the quality of these tests. For instance, other goodness-of-fit tests that consider subdivisions of the plot area into more than two parts, such as the chi-squared test might be suitable and more powerful and a comparison of the different approaches will be necessary. This should include a more detailed simulation study, also considering alternatives other than a linear trend. Finally, inhomogeneity may be more diffuse than a mere trend in x- or y-direction; tests that account for this type of inhomogeneity have yet to be developed and assessed in more detailed power studies.

There is also scope for further work in the context of the CUSUM approach. Other second order summary statistics may be considered and their performance assessed and compared to the performance yielded with the L-function described here. Similarly, the CUSUM approach may be further extended and applied to multivariate summary statistics, such as the paired L-functions. This would be particularly useful in a dataset with a very large number of potential interactions, such that interactions that are neither attractive nor inhibitive may be filtered out.

7.2.2 Multivariate methods for spatial point patterns

This thesis presents the first attempt to develop multivariate methods for spatial point pattern data. Naturally, there exists a large number of aspects that may be improved on and that have not been considered here. More

specifically, there is both scope for improving the FPCA approach taken here and for developing other multivariate methods based on functional data analysis methods.

This thesis focused on grouping the species based on their second order summary statistics, in particular the L-function and the pair correlation function. Clearly, other summary statistics, such as the K- and the J-function (van Lieshout and Baddeley 1999), may be applied in this context and their performance compared to the results obtained here. Similarly, the methods may be applied to the estimated first order summary statistics, i.e. the intensity surface, to yield a grouping of the species based on their first order characteristics or based on both the first and the second order characteristics. This would require a two-dimensional functional representation of the intensity surface through splines (Ramsay and Silverman 1997). Methods derived from shape analysis might also be usefully applied here (Dryden and Mardia 1998). Extensions to higher order summary statistics and pair-wise summary statistics may be applied in order to yield groupings based on higher order spatial characteristics as well as inter-species interaction.

Furthermore, the B-spline smoothing methods applied here are just one possibility of smoothing the summary statistics. More refined smoothing methods, such as other spline methods or wavelets which better take local features into account, may be more appropriate. This might be particularly advantageous in the context of inhomogeneous patterns, where the estimation bias at larger distances has caused problems. Note that a potential improved estimation of the intensity surface might improve the performance of the PCA.

As noted, other types of multivariate methods with different aims than the grouping of the species done by PCA may also be developed. For example, functional discriminant analysis (Ramsay and Silverman 1997) may be applied to the summary statistics. This may be used to test whether different pre-defined groups of species, such as families or species with different regeneration strategies, can be distinguished on the basis of their spatial pattern.

An FPCA approach based on several functions, i.e. several summary statistics for each species, may also be applied. This may be a particularly suitable approach when pair-wise second order summary statistics are to be incorporated since there will be a pair-wise interaction function for each species with all the other species.

Similarly, if marks are available a mixed PCA with FPCA on the spatial summary statistics pattern and a PCA on the marks may yield insight in the grouping structure of spatial behaviour combined with other biological information.

7.2.3 Modelling

The hierarchical modelling approach described in this thesis models the seeder species conditional on the location of the resprouter plants. This assumes that the seeders are independent of each other, i.e. no seeder-seeder interactions have been taken into account. However, it is likely that these interactions exist and a more advanced model would take these into account. Similarly, the spatial pattern formed by the resprouting species, and more specifically, the resprouter-resprouter interactions have not been modelled.

These are, however, likely to play an important role for community dynamics given that they have been established in the same locations for a long time.

It clearly may be considered a very appropriate strategy to use biological background information to simplify the complexity in a data set with a potentially large number of interactions. Nevertheless, in a more general context, e.g. in view of the rainforest data set mentioned in Chapter 2, a more general model would have to be constructed. This may involve introducing an asymmetric interaction structure where the hierarchical structure is not known. This may only be done if further information, e.g. marks, are available since otherwise the direction of the interaction cannot be identified. To this end, a model would have to be constructed which incorporates marks as well as data on the environmental conditions at the location of each of the individuals. Currently, work is being done in the context of the rainforest data where complex Cox models are being developed. There, one of the main challenges consists of jointly modelling the spatial variability of soil variables and the spatial point pattern. The soil variables cannot be modelled by a spatial point process but are data of a geostatistical nature as they are continuous in nature but may be only discretely measured. To date little work has been done on models that incorporate both geostatistical and point pattern data.

In addition, as data sets become available that have been collected repeatedly over time, spatial point process modelling will have to be extended to incorporate a temporal dimension. This is particularly useful in the context of plant communities since community dynamics evolve over time and since the pattern observed at a particular point in time is highly dependent

on the spatial structure and dynamics at previous points in time.

Furthermore, since no data on environmental conditions in the plants' microhabitats are available in the current data set, some questions regarding the validity of either the niche theory or the random drift theory could not be clarified. With an extension of the methodology to more complex data incorporating the rainforest data sets there is ample scope for this.

An interdisciplinary working group as mentioned in Section 7.1.3 has been formed to develop further spatial point process methodology suitable in the context of plant communities and to make the methodologies accessible to a wider group of non-specialist researchers. Within the working group a number of subgroups generalise the methods developed in this thesis. This includes the development of more general models than the hierarchical model described in Chapter 6, which incorporate soil variables and plant sizes as well as pattern development over time, as noted. The idea of multivariate methods for spatial point process methods based on functional data analysis is also being explored further e.g., by comparing the spatial patterns of species in different age groups.

Appendix A

List of plant species

species	species number	abundance	regeneration method
Actinotus leucocephalus	1	1	annual
Adenathos cygnorum	2	21	seeder
Alexgeorgea nitens	3	977	resprouter
Allocasuarina humilis	4	2	resprouter
Andersonia heterophylla	5	686	seeder
Astroloma xerophyllum	6	91	seeder
Baeckea grandiflora	7	1	resprouter
Banksia attenuata	8	26	resprouter
Banksia grandis	10	1	resprouter
Banksia ilicifolia	11	3	resprouter
Banksia menziesii	12	26	resprouter
Boronia ramosa	14	30	seeder
Bossieae eriocarpa	15	103	resprouter
Calitrix sapphirina	16	3	seeder
Conospermum acerosum	17	2	resprouter
Conospermum crassinervium	18	266	seeder
Conostephium pendulum	19	61	resprouter
Conostylis canscens	13	149	resprouter
Conostylis juncea	20	28	resprouter
Conostylis teretifolia	21	2	resprouter
Dampiera linearis	22	24	resprouter
Dasypogon bromeliifolius	23	167	resprouter
Daviessia aff. *incrassata*	24	1	resprouter
Eremaea astrocarpa	25	207	resprouter
Eriostemon spicatus	26	65	resprouter/ seeder
Eucalyptus todtiana	27	1	resprouter
Gompholobium tomentosum	28	9	seeder
Hakea ruscifolia	29	1	resprouter
Hakea varia	30	1	resprouter
Hensmania turbinata	31	15	resprouter
Hibbertia crassifolia	32	96	resprouter
Hibbertia hypericoides	33	148	resprouter
Hibbertia sp.	34	134	resprouter
Hibbertia subvaginata	35	5	resprouter

Table A.1: List of plant species contained in the Cooljarloo data set, part I

species	species number	abundance	regeneration method
Hypocalymma xantopethalum	36	96	resprouter
Isopogon linearis	37	68	resprouter
Jacksonia floribunda	38	124	resprouter
Jacksonia spinosa	39	8	seeder
Johnsonia pubescens	41	1	seeder
Chordifex sinuosus	42	154	resprouter
Kingia australis	43	1	resprouter
Lepidosperma angustatum	44	22	resprouter
Lepidosperma tenue	45	61	resprouter
Leporella fimbriata	40	1	resprouter
Leptospermum spinescens	46	8	resprouter
Leucopogon conostephioides	47	657	seeder
Leucopogon striatus	48	251	seeder
Lomandra sp.	49	304	resprouter
Lyginia barbata	50	299	resprouter
Melaleuca scabra	51	377	resprouter/ seeder
Mesomelaena pseudostygia	52	16	resprouter
Montaxis grandifloria	53	1	resprouter
Patersonia occidentalis	54	79	resprouter
Petrophile linearis	56	3	resprouter
Petrophile seminuda	55	14	seeder
Phlebocarya philifolia	57	207	resprouter
Pimelea sulphurea	58	12	resprouter
Platysace juncea	59	22	resprouter
Platysace xerophila	60	1	resprouter
Scholtzia involucrata	61	170	resprouter
Stirlingia latifolia	62	15	resprouter/ seeder
Stylidium brunonianum	63	3	seeder
Stylidium crossocephalum	64	27	seeder
Thomasia grandifloria	65	1	resprouter
Thysonatus sparteus	66	3	resprouter
Xanthorrhoea drummondii	67	2	resprouter
Xanthosia huegelii	68	14	resprouter

Table A.2: List of plant species contained in the Cooljarloo data set, part II

Appendix B

Plots and summary statistics for all patterns

B.1 Plots and first order summary statistics for all species

B.1.1 Patterns and intensity surfaces

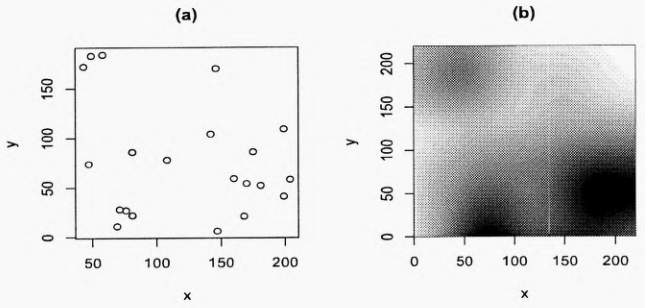

Figure B.1: Point pattern (a) and estimated intensity (b) for *Adenathos cygonum*

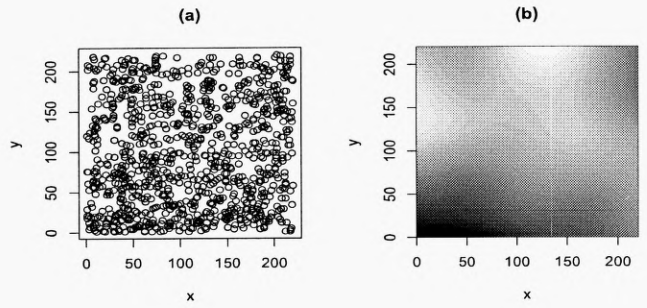

Figure B.2: Point pattern (a) and estimated intensity (b) for *Alexgeorgia nitens*

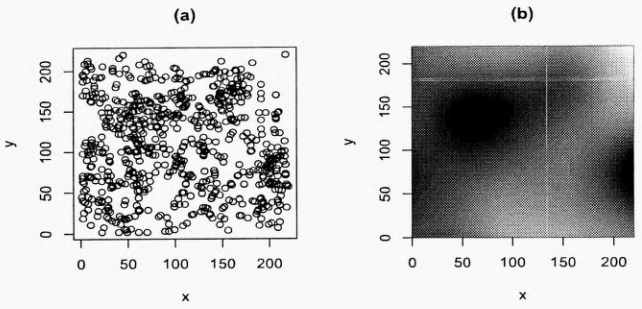

Figure B.3: Point pattern (a) and estimated intensity (b) for *Andersonia heterophylla*

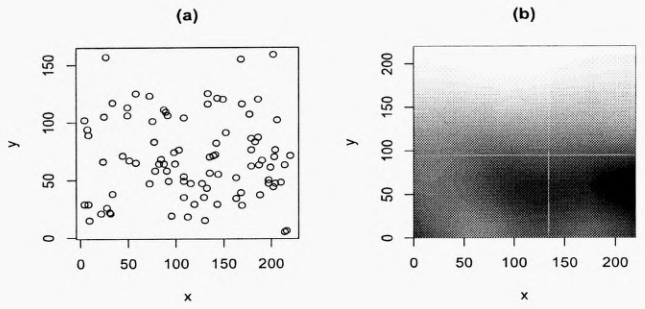

Figure B.4: Point pattern (a) and estimated intensity (b) for *Astroloma xerophyllum*

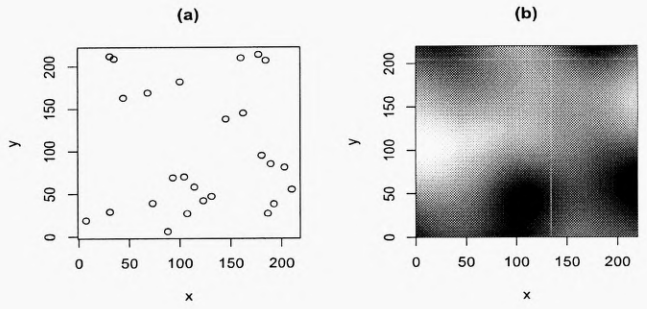

Figure B.5: Point pattern (a) and estimated intensity (b) for *Banksia attenuata*

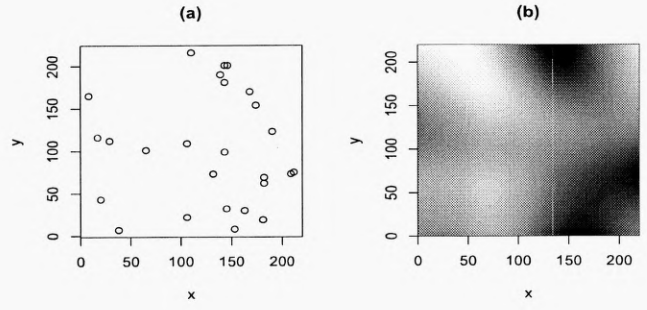

Figure B.6: Point pattern (a) and estimated intensity (b) for *Banksia menziesii*

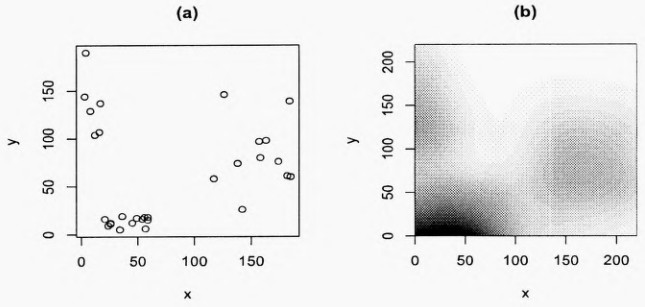

Figure B.7: Point pattern (a) and estimated intensity (b) for *Boronia ramosa*

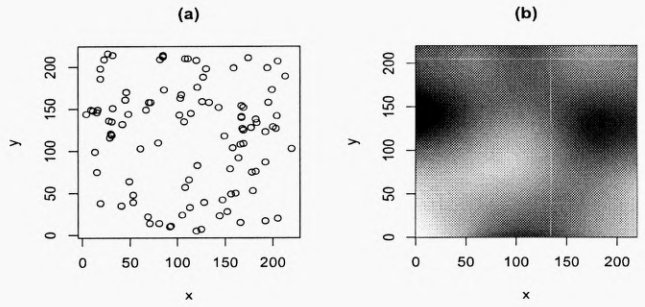

Figure B.8: Point pattern (a) and estimated intensity (b) for *Bossieae eriocarpa*

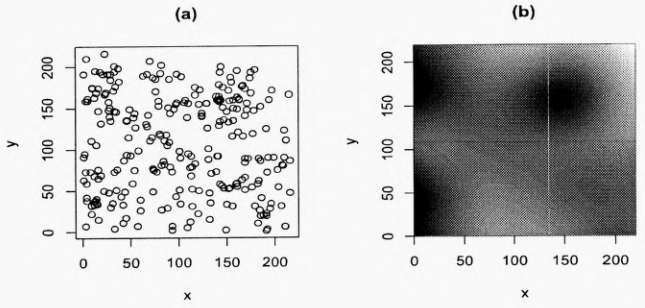

Figure B.9: Point pattern (a) and estimated intensity (b) for *Conospermum crassinervium*

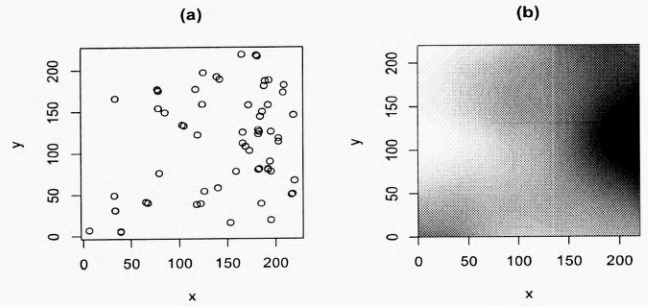

Figure B.10: Point pattern (a) and estimated intensity (b) for *Conostephium pendulum*

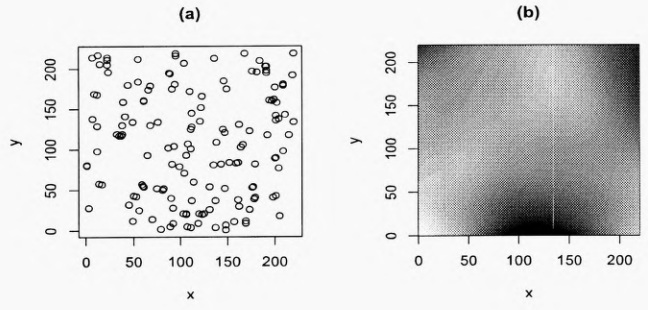

Figure B.11: Point pattern (a) and estimated intensity (b) for *Conostylis canscens*

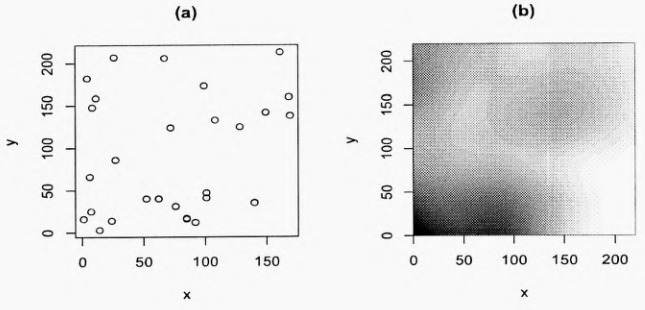

Figure B.12: Point pattern (a) and estimated intensity (b) for *Conostylis juncea*

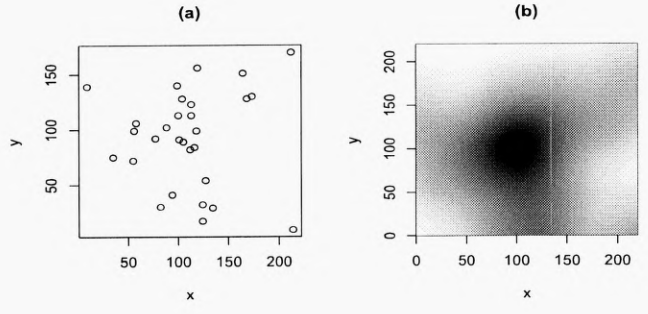

Figure B.13: Point pattern (a) and estimated intensity (b) for *Dampiera linearis*

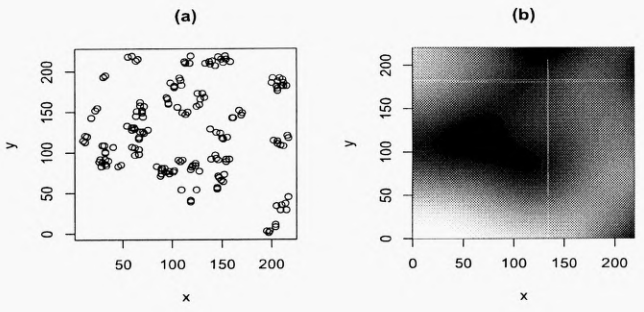

Figure B.14: Point pattern (a) and estimated intensity (b) for *Dasypogon bromeliifolius*

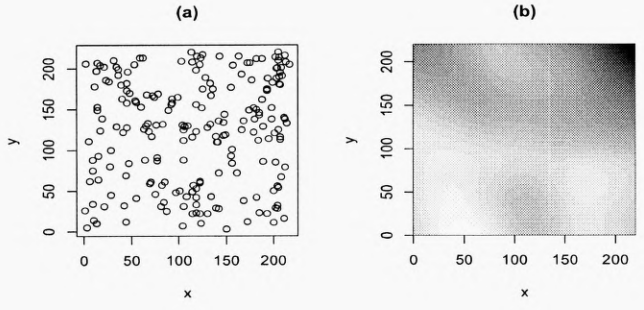

Figure B.15: Point pattern (a) and estimated intensity (b) for *Eremaea astrocarpa*

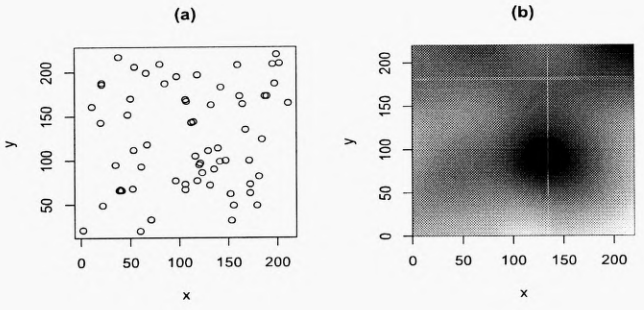

Figure B.16: Point pattern (a) and estimated intensity (b) for *Eriostemon spicatus*

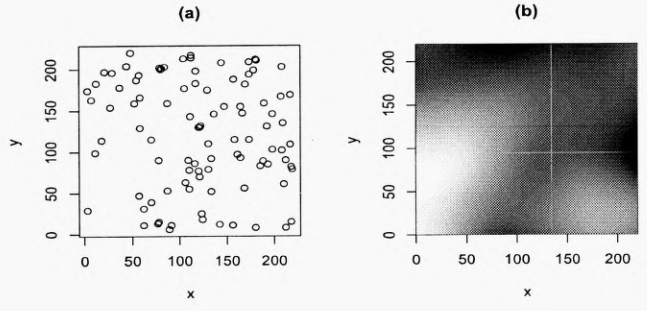

Figure B.17: Point pattern (a) and estimated intensity (b) for *Hibbertia crassifolia*

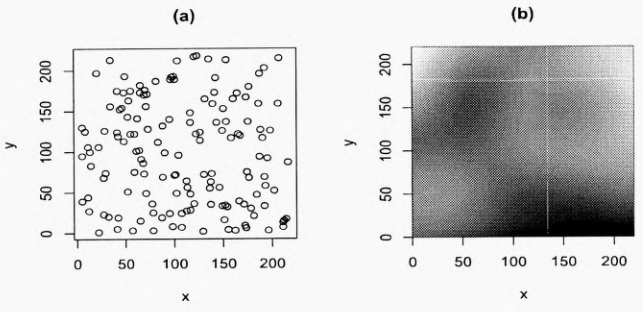

Figure B.18: Point pattern (a) and estimated intensity (b) for *Hibbertia hypericoides*

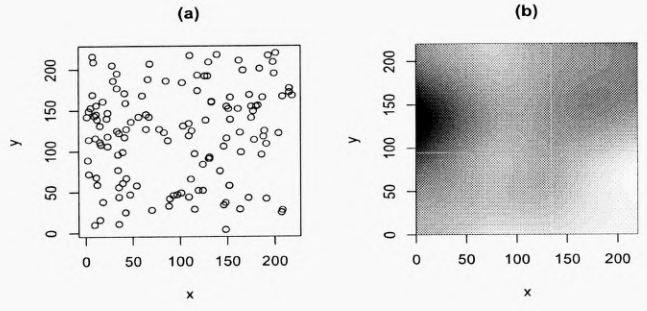

Figure B.19: Point pattern (a) and estimated intensity (b) for *Hibbertia subvaginata*

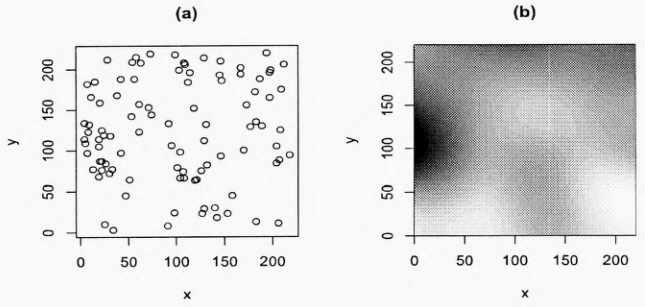

Figure B.20: Point pattern (a) and estimated intensity (b) for *Hypocalymna xanthopetalum*

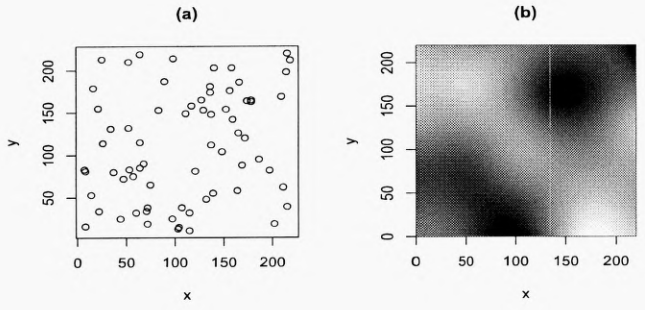

Figure B.21: Point pattern (a) and estimated intensity (b) for *Isopogon linearis*

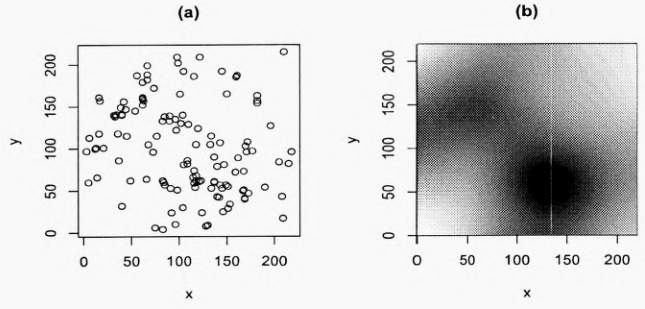

Figure B.22: Point pattern (a) and estimated intensity (b) for *Jacksonia floribunda*

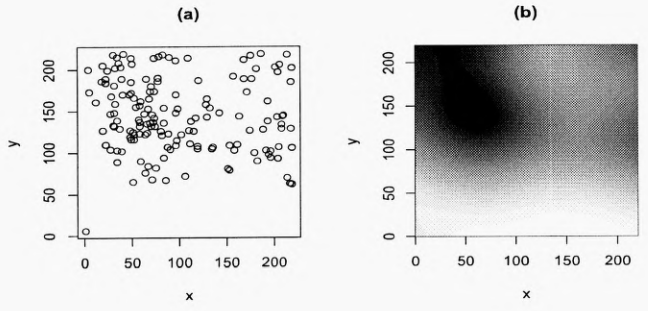

Figure B.23: Point pattern (a) and estimated intensity (b) for *"Restio sinuousus"*

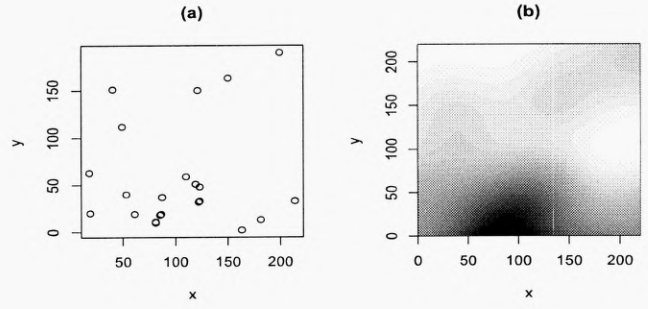

Figure B.24: Point pattern (a) and estimated intensity (b) for *Lepidosperma angistatum*

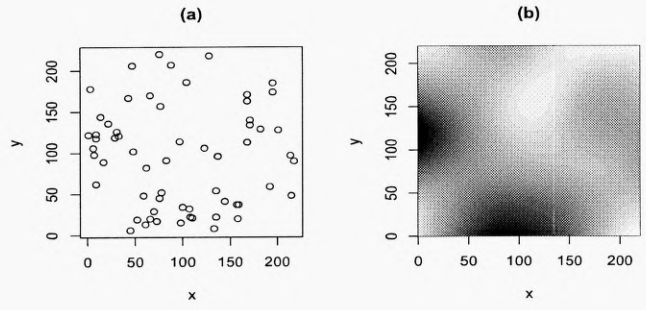

Figure B.25: Point pattern (a) and estimated intensity (b) for *Lepidosperma tenue*

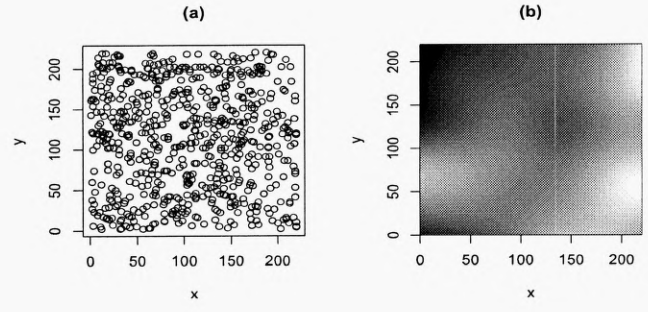

Figure B.26: Point pattern (a) and estimated intensity (b) for *Leucopogon conostephioides*

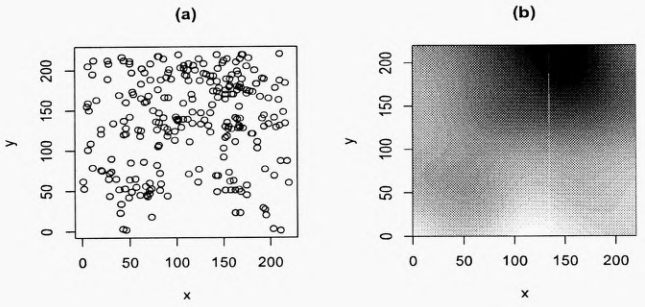

Figure B.27: Point pattern (a) and estimated intensity (b) for *Leucopogon striatus*

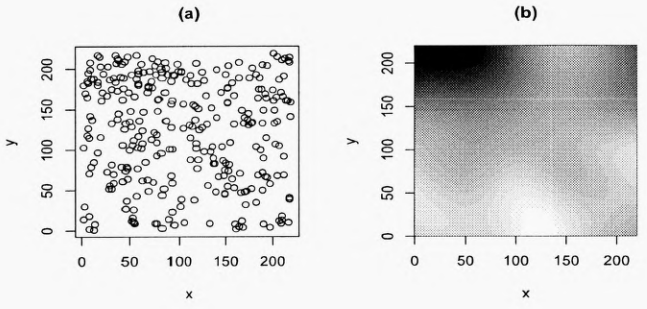

Figure B.28: Point pattern (a) and estimated intensity (b) for *Lomandra sp.*

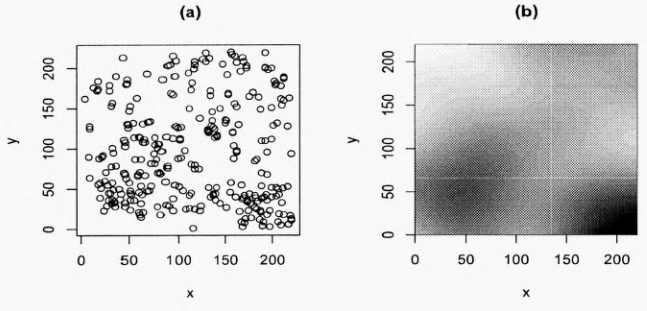

Figure B.29: Point pattern (a) and estimated intensity (b) for *Lyginia barbata*

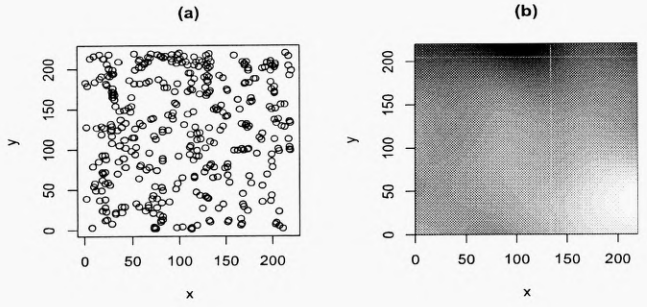

Figure B.30: Point pattern (a) and estimated intensity (b) for *Melaleuca scabra*

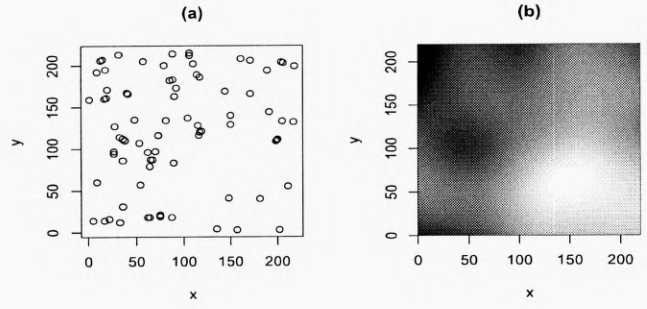

Figure B.31: Point pattern (a) and estimated intensity (b) for *Patersonia occidentalis*

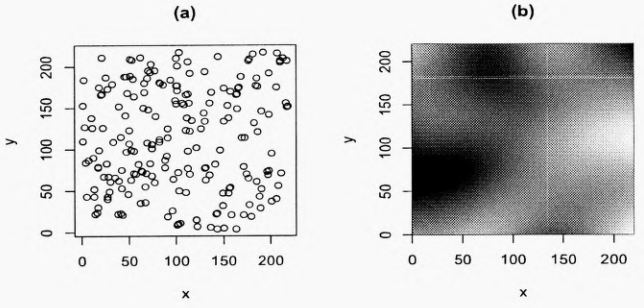

Figure B.32: Point pattern (a) and estimated intensity (b) for *Phlebocarya filifolia*

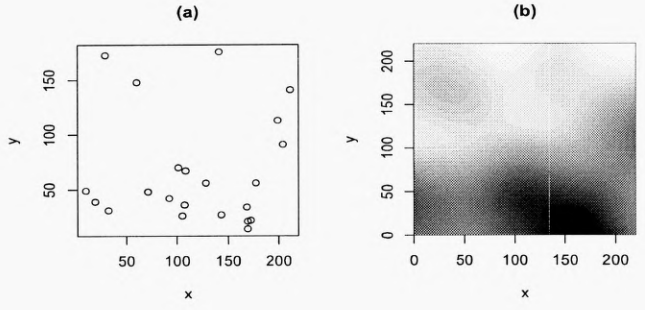

Figure B.33: Point pattern (a) and estimated intensity (b) for *Platysace juncea*

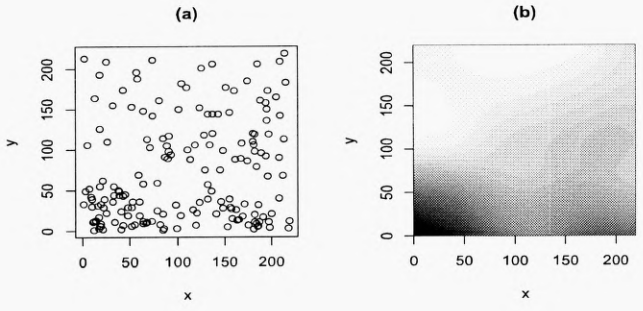

Figure B.34: Point pattern (a) and estimated intensity (b) for *Scholtzia involucrata*

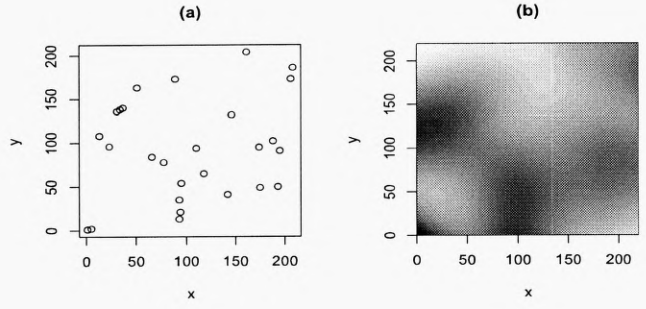

Figure B.35: Point pattern (a) and estimated intensity (b) for *Stylidium crossocephalum*

B.1.2 Results of test for inhomogeneity

species	F-test	Permutation test	Kolmogorov-Smirnov x-direction y-direction
Adenanthos cygnorum	$p = 0.063$	$p = 0.000$	$p = 0.352$ $p = 0.019$
Alexgeorgia nitens	$p = 0.89$	$p = 0.970$	$p = 0.186$ $p = 0.000$
Andersonia heterophylla	$p = 0.967$	$p = 0.000$	$p = 0.004$ $p = 0.003$
Astroloma xerophyllum	$p = 0.001$	$p = 0.000$	$p = 0.36$ $p = 0.000$
Banksia attenuata	$p = 0.734$	$p = 0.71$	$p = 0.72$ $p = 0.27$
Banksia menziesii	$p = 0.453$	$p = 0.23$	$p = 0.058$ $p = 0.69$
Boronia ramosa	$p = 0.023$	$p = 0.13$	$p = 0.0004$ $p = 0.0009$
Bossiaea eriocarpa	$p = 0.497$	$p = 0.11$	$p = 0.90$ $p = 0.0375$
Conospermum crassinervium	$p = 0.274$	$p = 0.23$	$p = 0.08$ $p = 0.0487$
Conostephium pendulum	$p = 0.730$	$p = 0.59$	$p = 0.000$ $p = 0.889$
Conostylis canescens	$p = 0.784$	$p = 0.83$	$p = 0.40$ $p = 0.42$
Conostylis juncea	$p = 0.882$	$p = 0.78$	$p = 0.0119$ $p = 0.049$
Dampiera linearis	$p = 0.047$	$p = 0.000$	$p = 0.11$ $p = 0.03$
Dasypogon bromeliifolius	$p = 0.000$	$p = 0.000$	$p = 0.02$ $p = 0.000$
Eremaea astrocarpa	$p = 0.770$	$p = 0.69$	$p = 0.11$ $p = 0.0004$
Eriostemon spicatus	$p = 0.03$	$p = 0.01$	$p = 0.57$ $p = 0.04$

Table B.1: p-values for three tests of inhomogeneity of patterns for all species

species	F-test	Permutation test	Kolmogorov-Smirnov x-direction y-direction
Hibbertia crassifolia	$p = 0.58$	$p = 0.57$	$p = 0.082$ $p = 0.19$
Hibbertia hypericoides	$p = 0.34$	$p = 0.27$	$p = 0.7022$ $p = 0.11$
Hibbertia sp.	$p = 0.422$	$p = 0.37$	$p = 0.0057$ $p = 0.0117$
Hypocalymma xanthopetalum	$p = 0.344$	$p = 0.44$	$p = 0.149$ $p = 0.0252$
Isopogon linearis	$p = 0.729$	$p = 0.51$	$p = 0.9624$ $p = 0.9398$
Jacksonia floribunda	$p = 0.982$	$p = 0.000$	$p = 0.0676$ $p = 0.0322$
Chordifex sinuosus	$p = 0.147$	$p = 0.09$	$p = 0.0005$ $p = 0.000$
Lepidosperma angistatum	$p = 0.29$	$p = 0.45$	$p = 0.2322$ $p = 0.000$
Lepidosperma tenue	$p = 0.65$	$p = 0.64$	$p = 0.315$ $p = 0.2098$
Leucopogon conostephioides	$p = 0.265$	$p = 0.41$	$p = 0.0234$ $p = 0.0103$
Leucopogon striatus	$p = 0.914$	$p = 0.000$	$p = 0.131$ $p = 0.000$
Lomandra sp.	$p = 0.532$	$p = 0.42$	$p = 0.493$ $p = 0.0002$
Lyginia barbata	$p = 0.144$	$p = 0.07$	$p = 0.207$ $p = 0.000$
Melaleuca scabra	$p = 0.533$	$p = 0.68$	$p = 0.038$ $p = 0.000$
Patersonia occidentalis	$p = 0.45$	$p = 0.62$	$p = 0.012$ $p = 0.07$
Phlebocarya filifolia	$p = 0.540$	$p = 0.38$	$p = 0.067$ $p = 0.67$
Platysace juncea	$p = 0.347$	$p = 0.51$	$p = 0.6875$ $p = 0.0013$
Scholtzia involucrata	$p = 0.989$	$p = 0.67$	$p = 0.45$ $p = 0.000$
Stylidium crossocephalum	$p = 0.844$	$p = 0.60$	$p = 0.7575$ $p = 0.3174$

Table B.2: p-values for three tests of inhomogeneity of patterns for all species

B.2 Second order summary statistics for all species

In the following we present the second order summary statistics for all species with more than 20 plants. When the inhomogeneity tests in section 4.1.2 indicated that the pattern was inhomgeneous the inhomgeneous L-function (Baddeley et al. 2000) was used, otherwise the homogeneous L-function as well as the inhomgeneous L-function (which should yield the same result) were applied.

The examples presented in chapter 4 are repeated here for the sake of completeness.

B.2.1 Plots of summary statistics

B.2.1.1 Homogeneous patterns

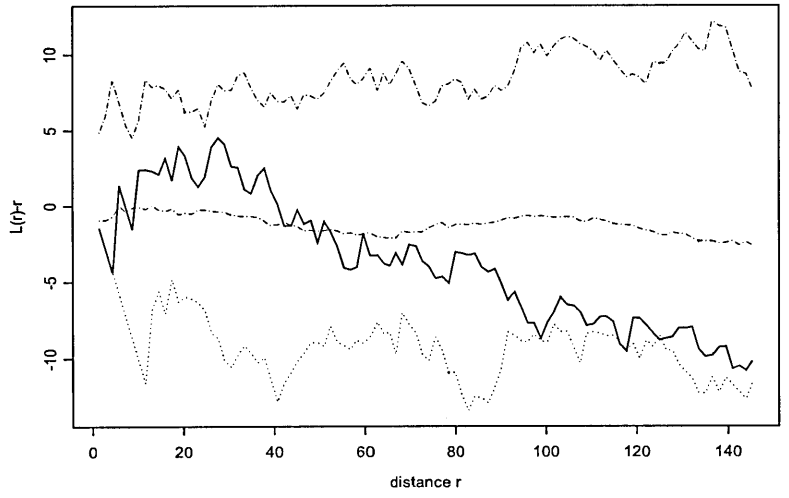

Figure B.36: Homogeneous L-function with simulation envelopes for *Banksia attenuata*

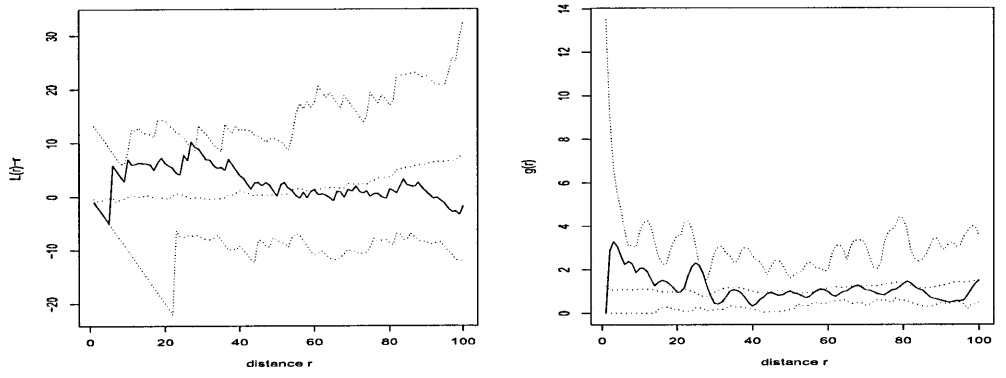

Figure B.37: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Banksia attenuata*

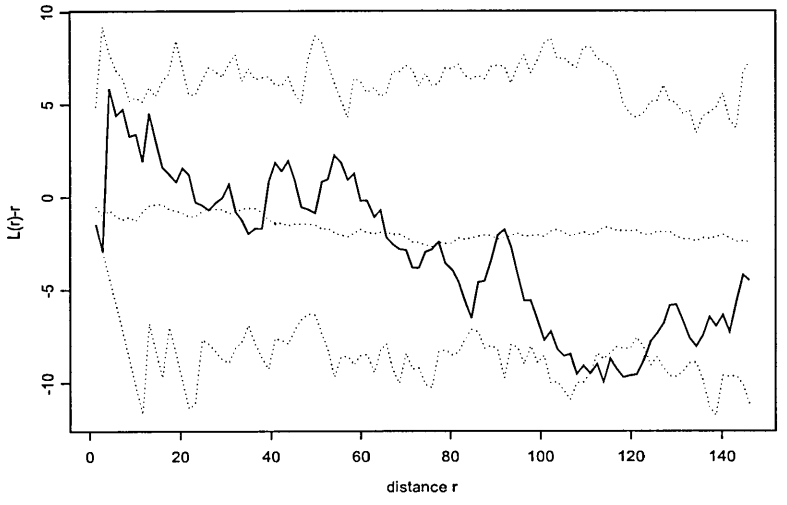

Figure B.38: Homogeneous L-function with simulation envelopes for *Banksia menziesii*

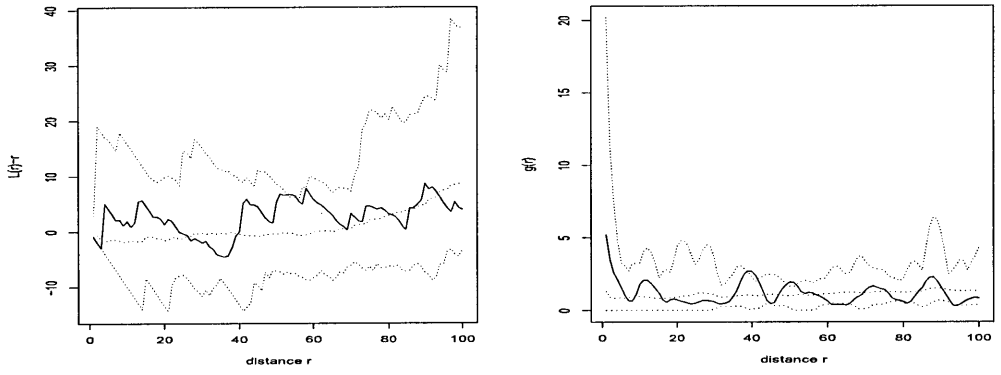

Figure B.39: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Banksia menziesii*

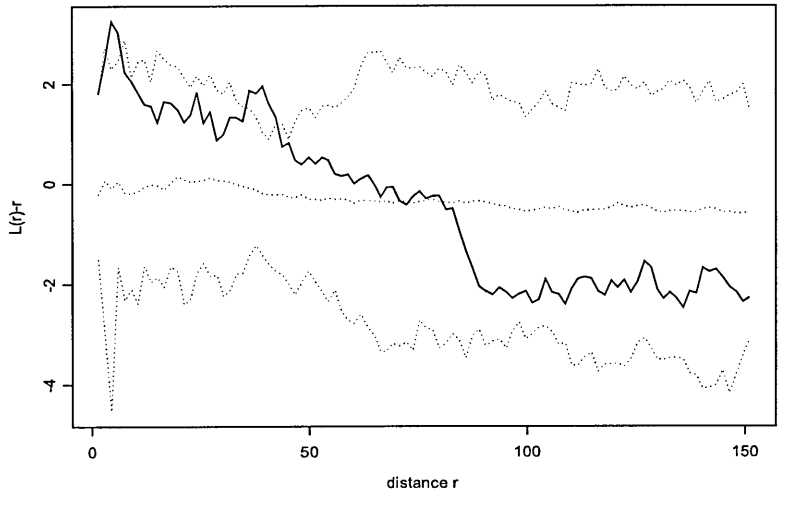

Figure B.40: Homogeneous L-function with simulation envelopes for *Bossiaea eriocarpa*

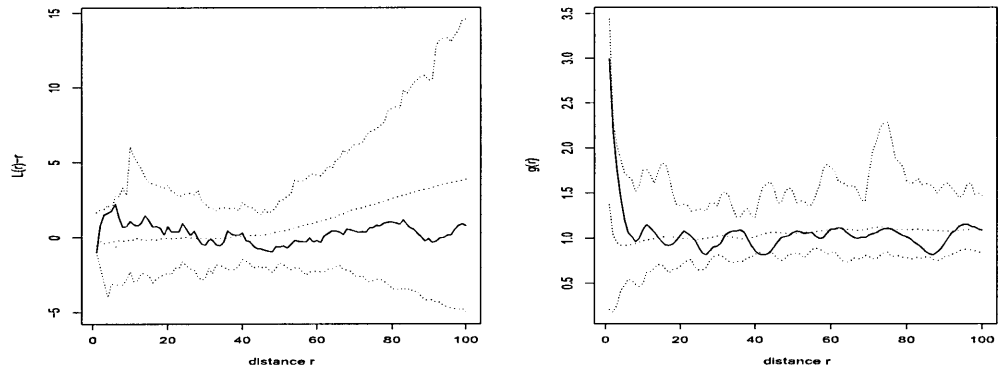

Figure B.41: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Bossiaea eriocarpa*

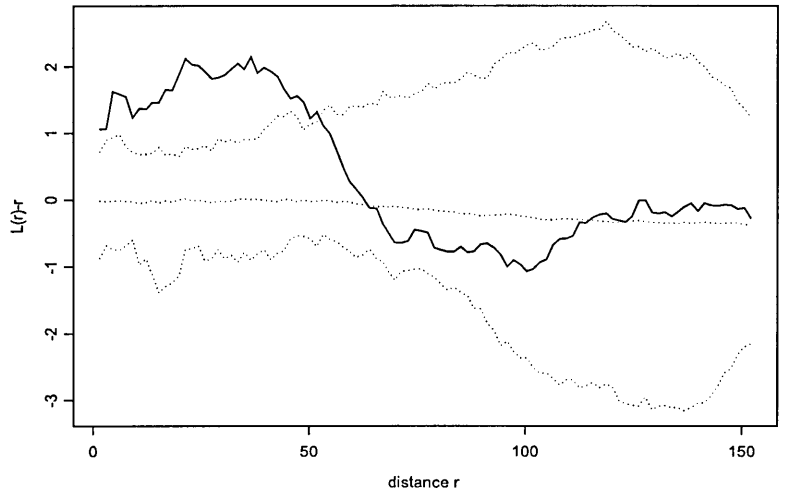

Figure B.42: Homogeneous L-function with simulation envelopes for *Conospermum crassinervium*

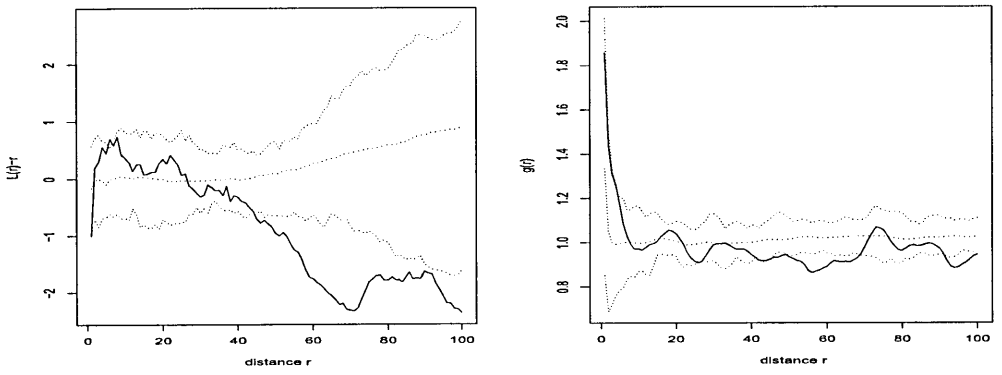

Figure B.43: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Conospermum crassinervium*

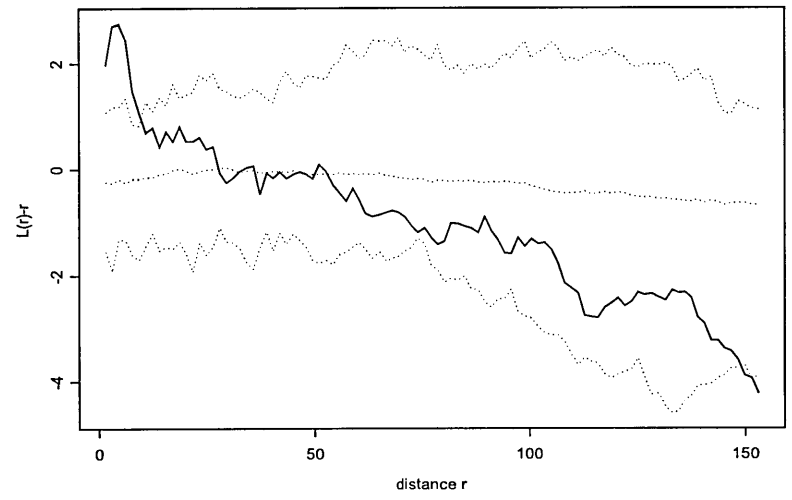

Figure B.44: Homogeneous L-function with simulation envelopes for *Conostylis canescens*

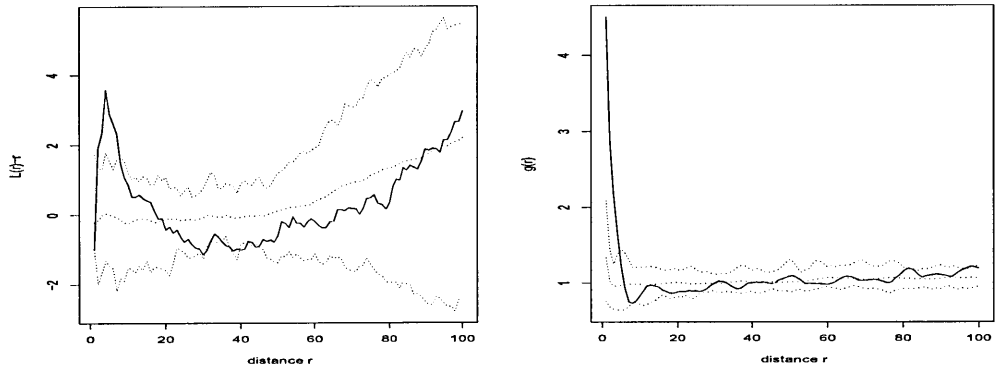

Figure B.45: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Conostylis canescens*

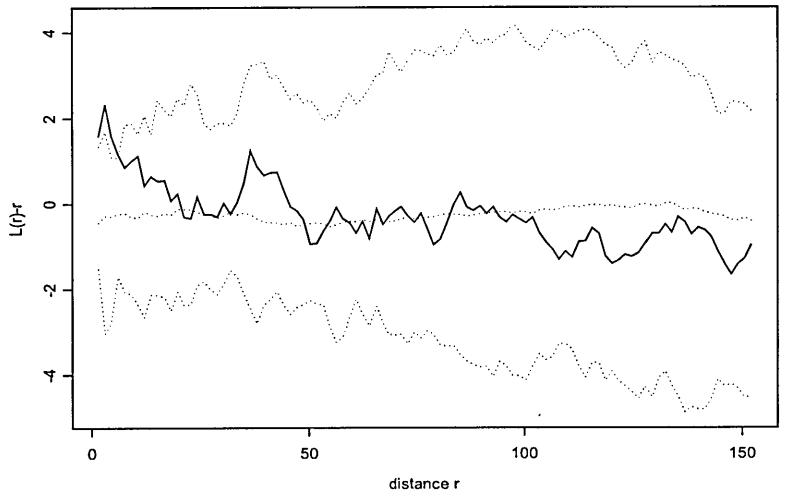

Figure B.46: Homogeneous L-function with simulation envelopes for *Hibbertia crassifolia*

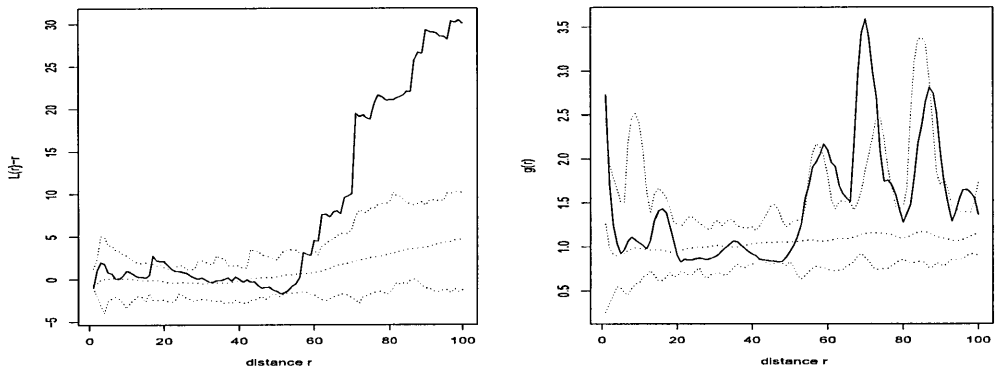

Figure B.47: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Hibbertia crassifolia*

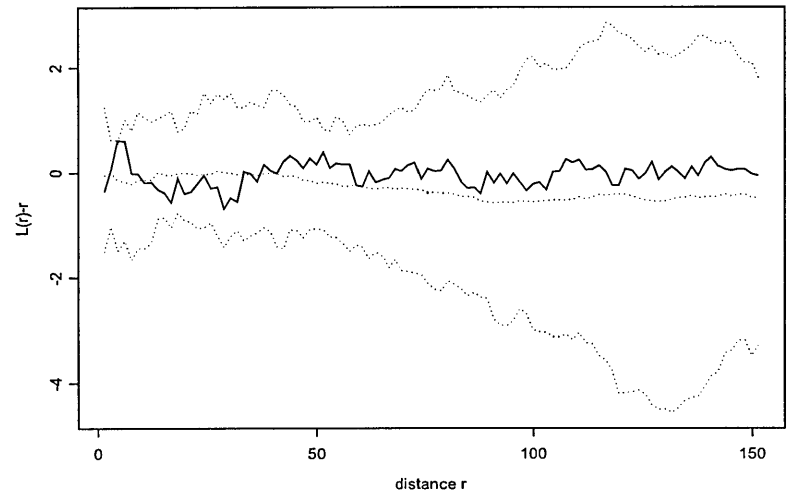

Figure B.48: Homogeneous L-function with simulation envelopes for *Hibbertia hypericoides*

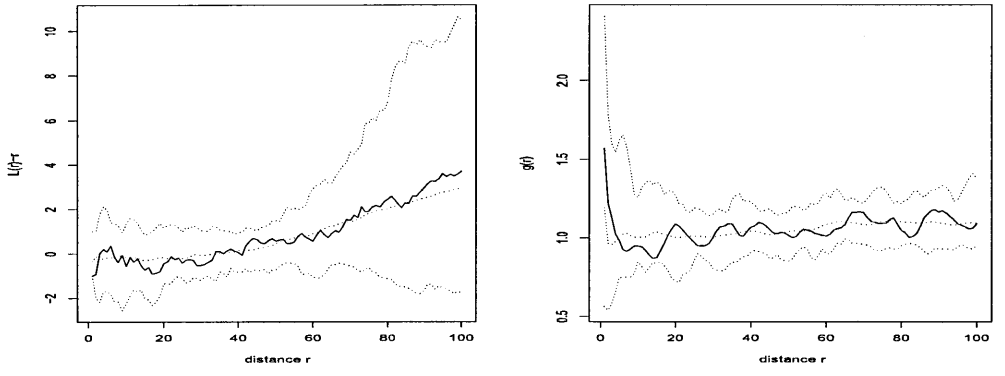

Figure B.49: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Hibbertia hypericoides*

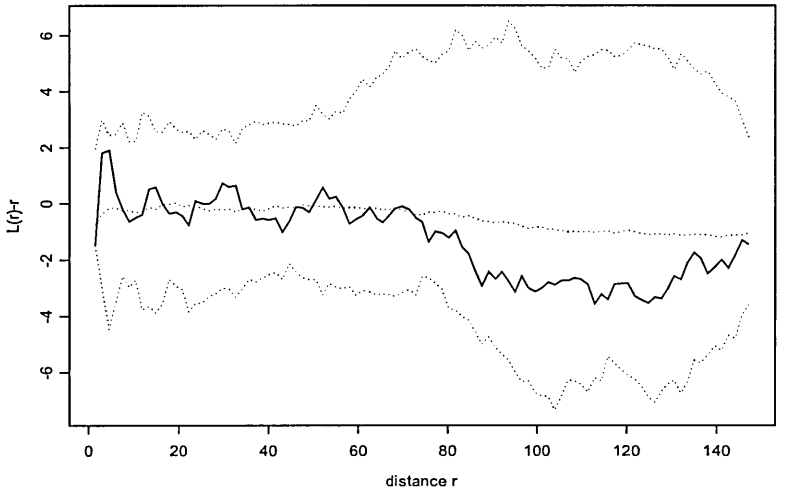

Figure B.50: Homogeneous L-function with simulation envelopes for *Isopogon linearis*

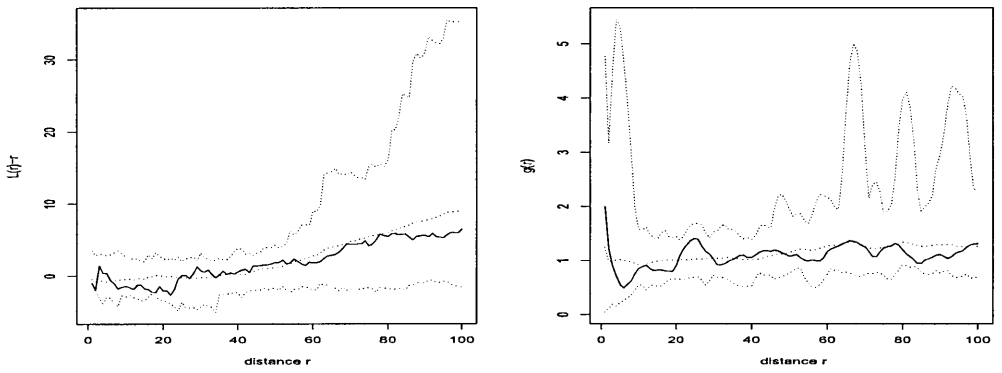

Figure B.51: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Isopogon linearis*

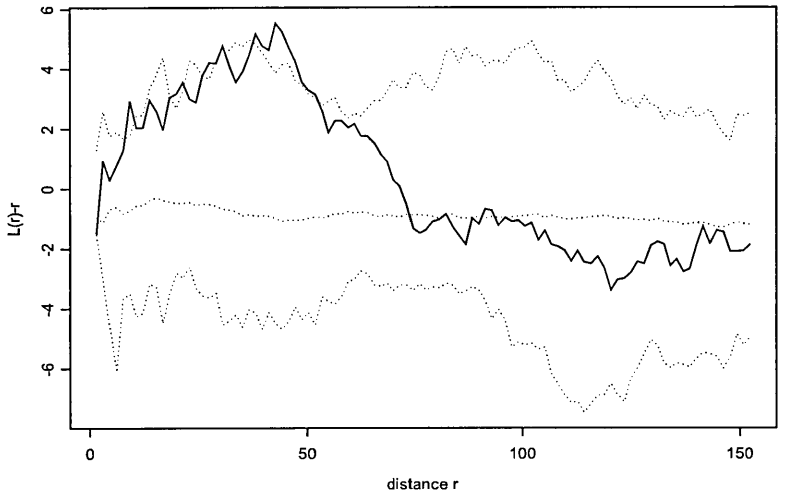

Figure B.52: Homogeneous L-function with simulation envelopes for *Lepidosperma tenue*

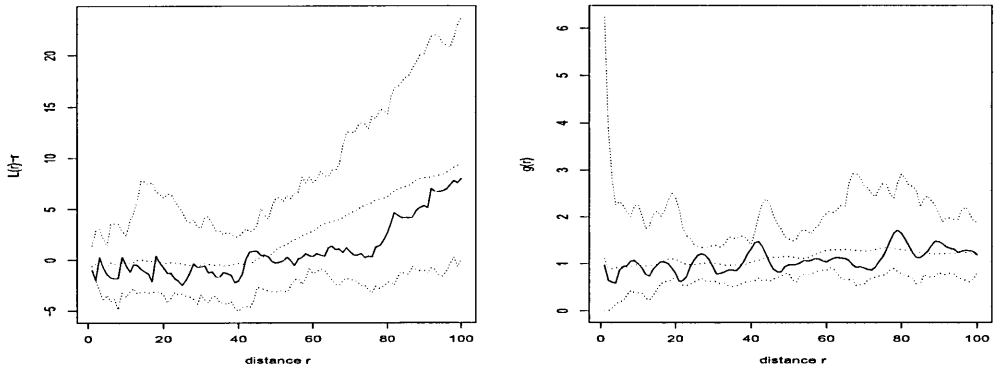

Figure B.53: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Lepidosperma tenue*

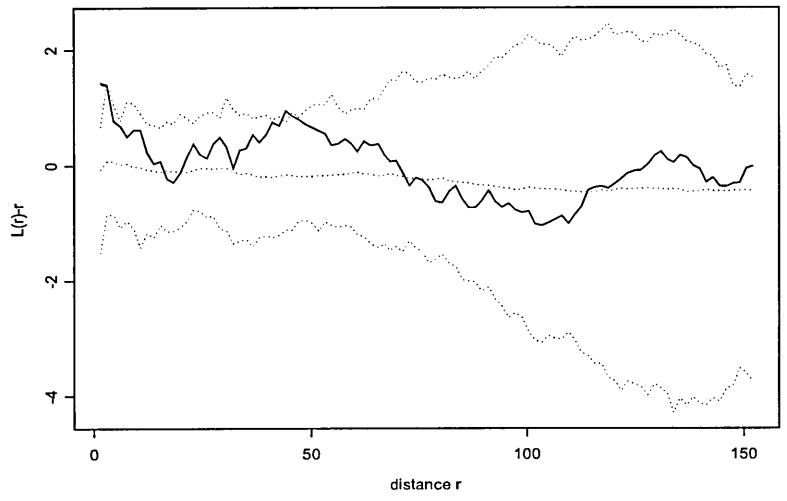

Figure B.54: Homogeneous L-function with simulation envelopes for *Phlebocarya filifolia*

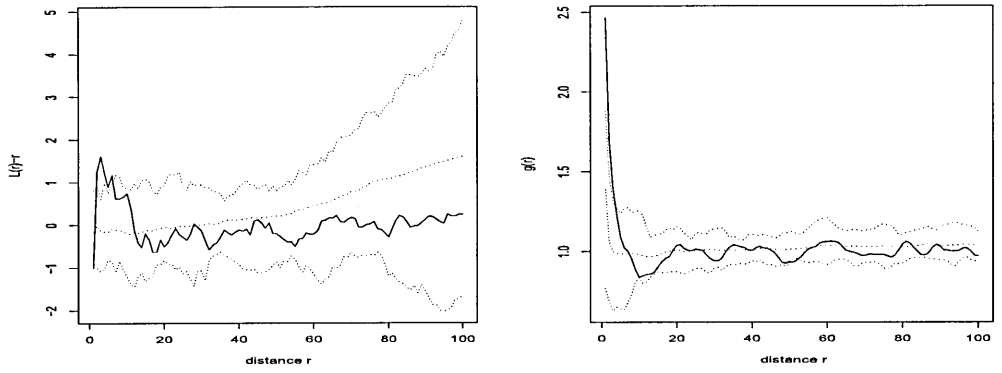

Figure B.55: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Phlebocarya filifolia*

B.2.1.2 Inhomogeneous patterns

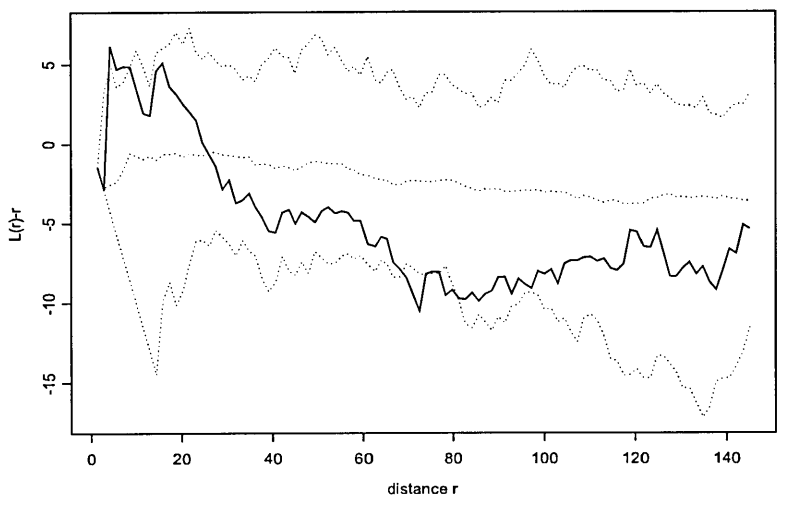

Figure B.56: Homogeneous L-function with simulation envelopes for *Stylidium crossocephalum*

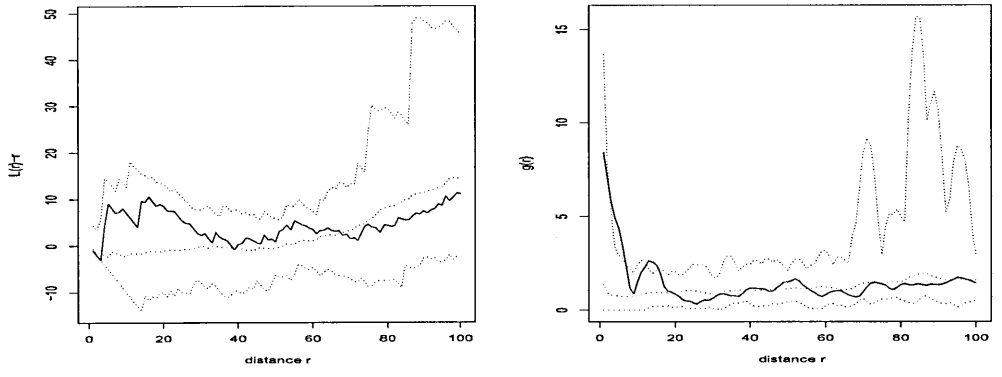

Figure B.57: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Stylidium crossocephalum*

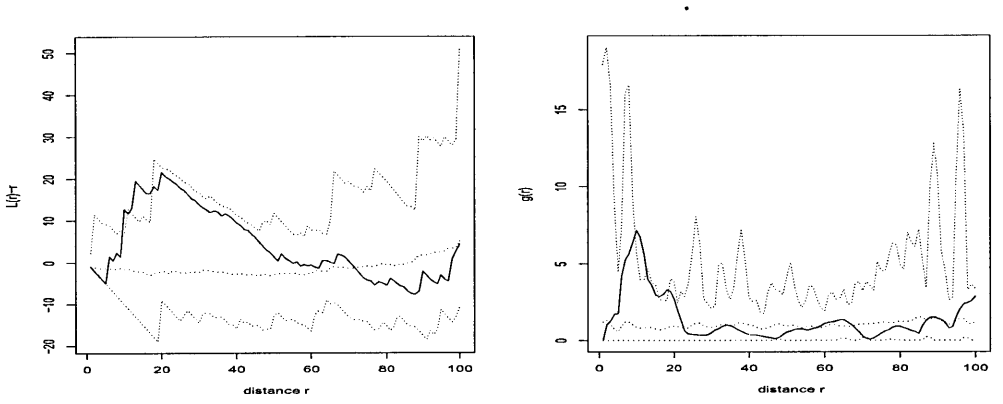

Figure B.58: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Adenathos cygnorum*

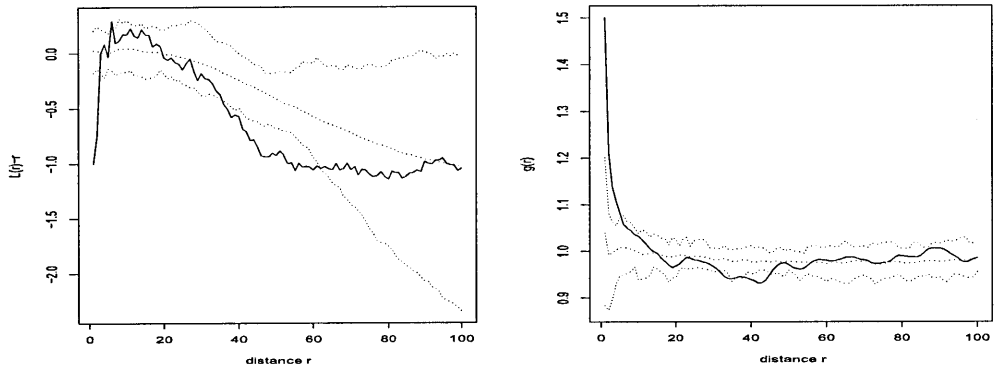

Figure B.59: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Alexgeogea nitens*

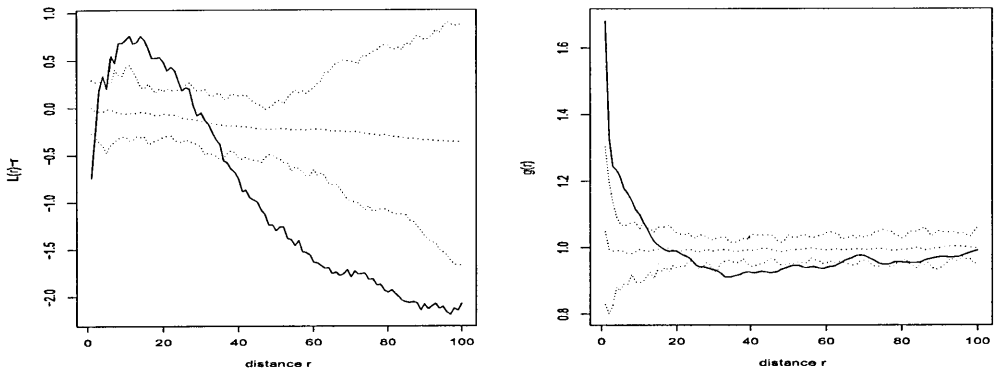

Figure B.60: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Andersonia heterophylla*

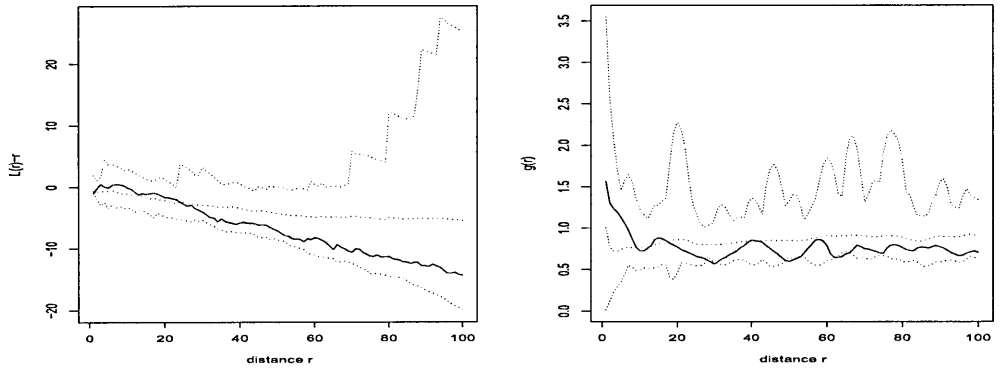

Figure B.61: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Astroloma Xerophyllum*

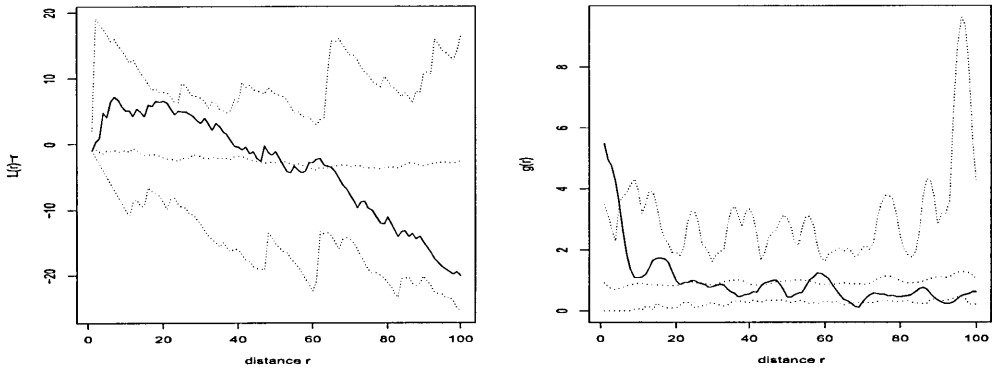

Figure B.62: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Boronia ramosa*

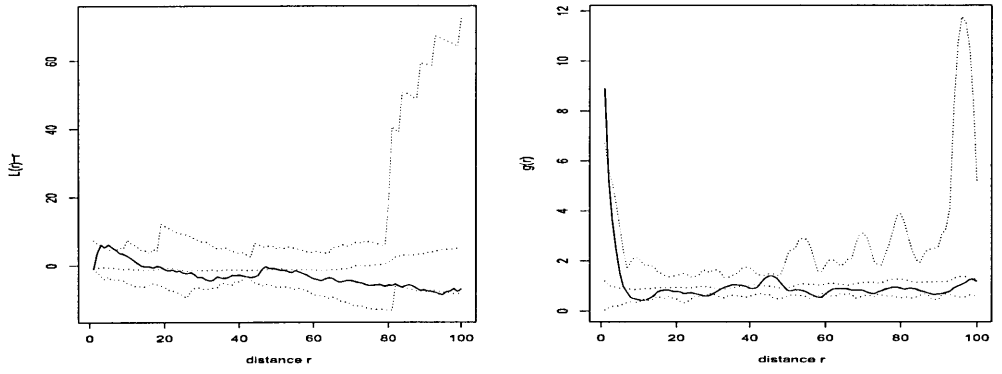

Figure B.63: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Conostephium pendulum*

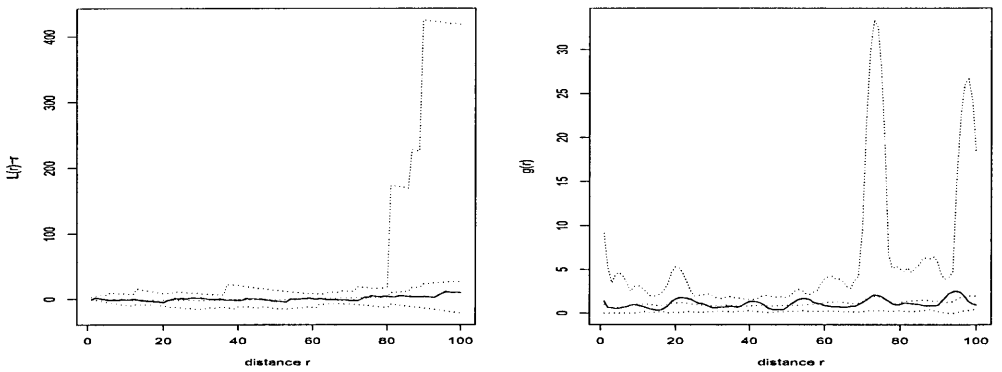

Figure B.64: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Conostylis juncea*

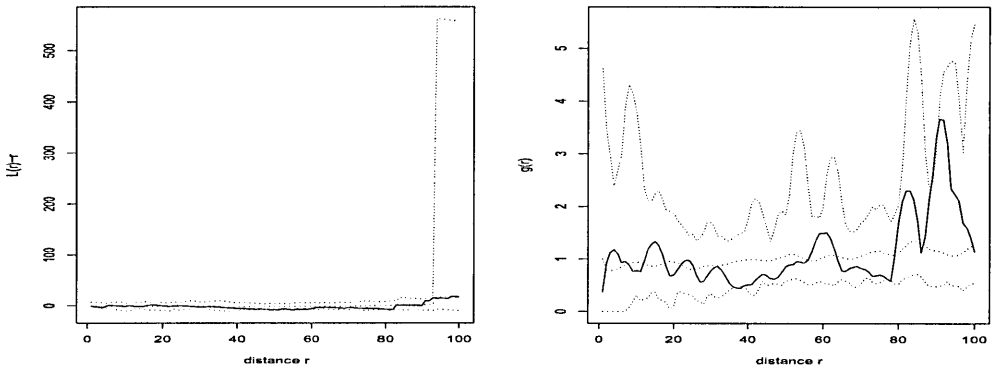

Figure B.65: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Dampiera linearis*

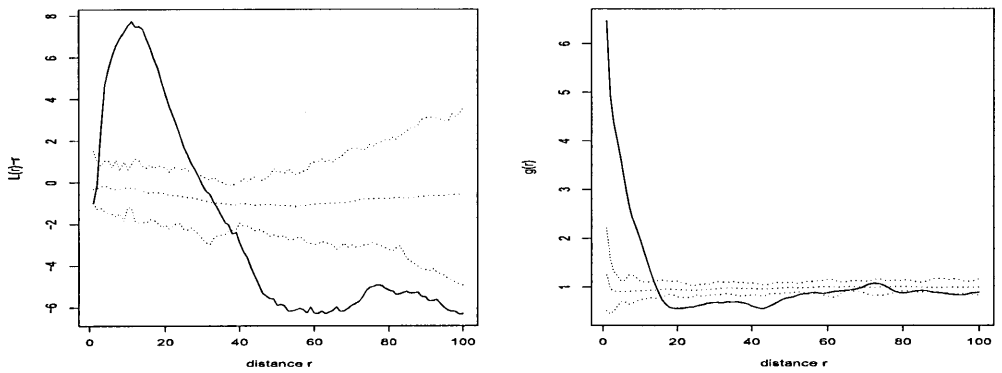

Figure B.66: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Dasypogon bromeliifolius*

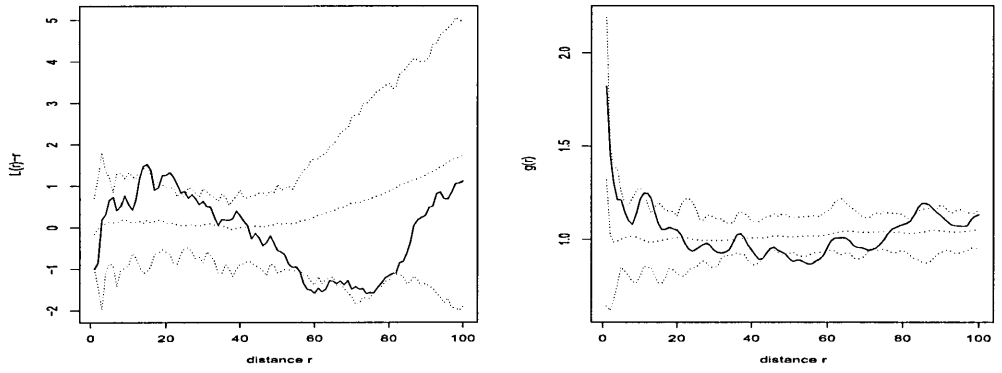

Figure B.67: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Eremaea astrocarpa*

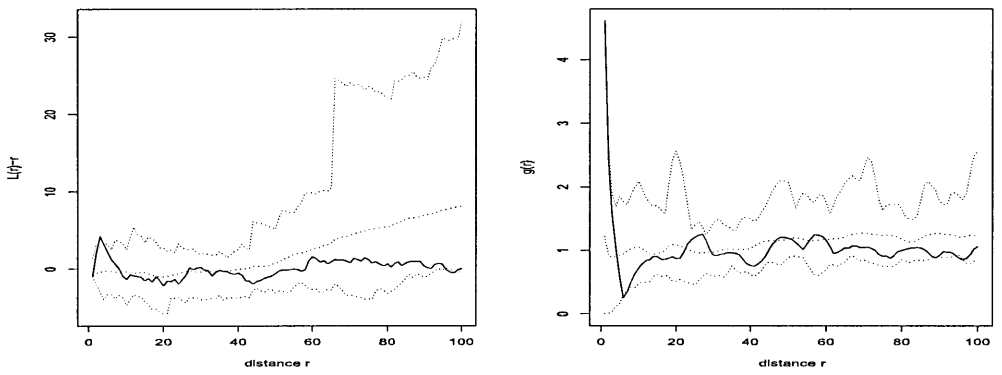

Figure B.68: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Eriostemon spicatus*

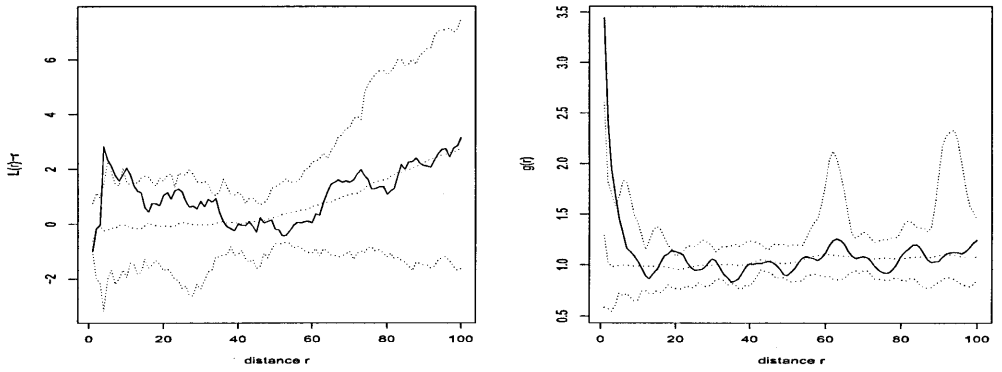

Figure B.69: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Hibbertia subvaginata*

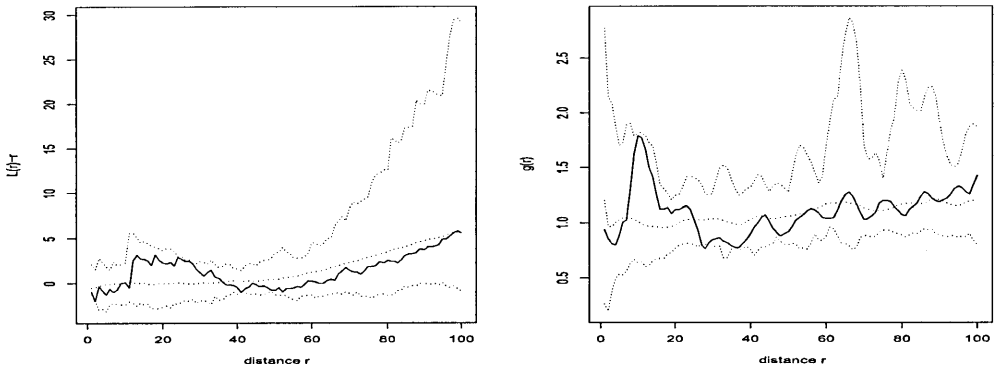

Figure B.70: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Hypocalymma xanthopetalum*

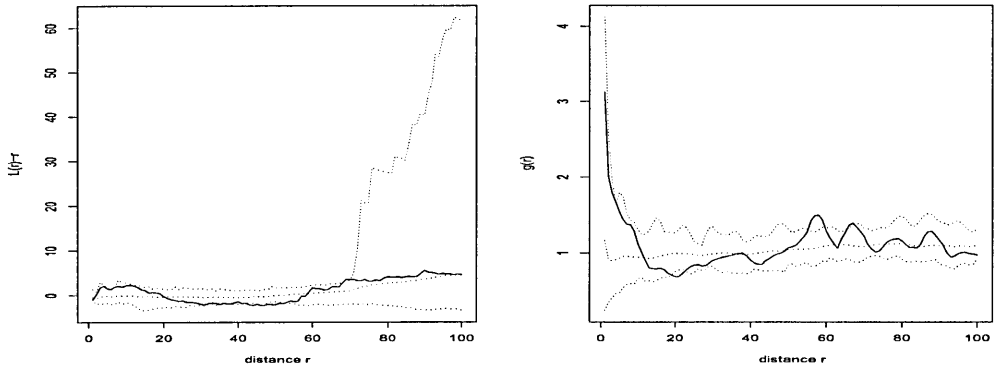

Figure B.71: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Jacksonia floribunda*

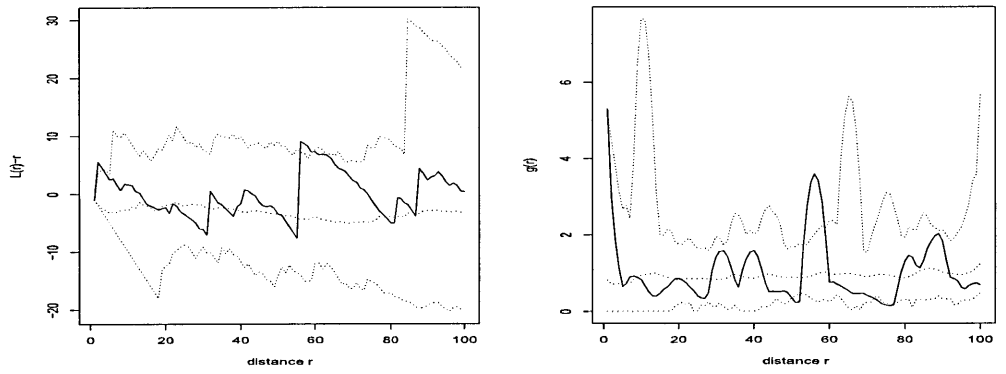

Figure B.72: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Lepidosperma angistatum*

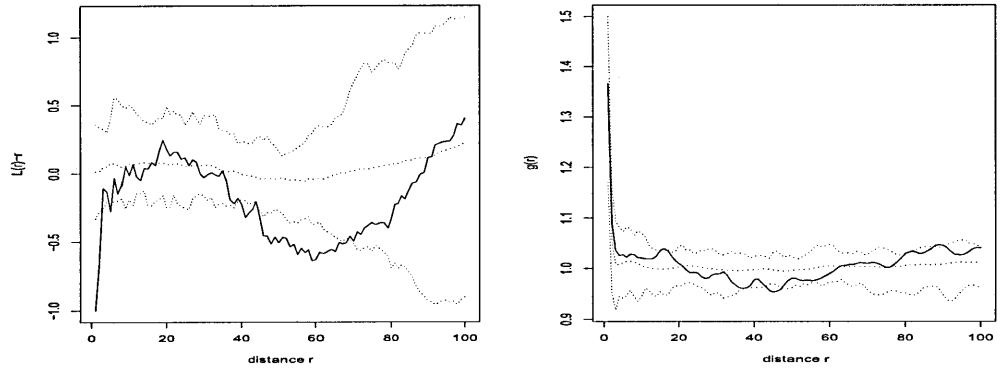

Figure B.73: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Leucopogon conostephioides*

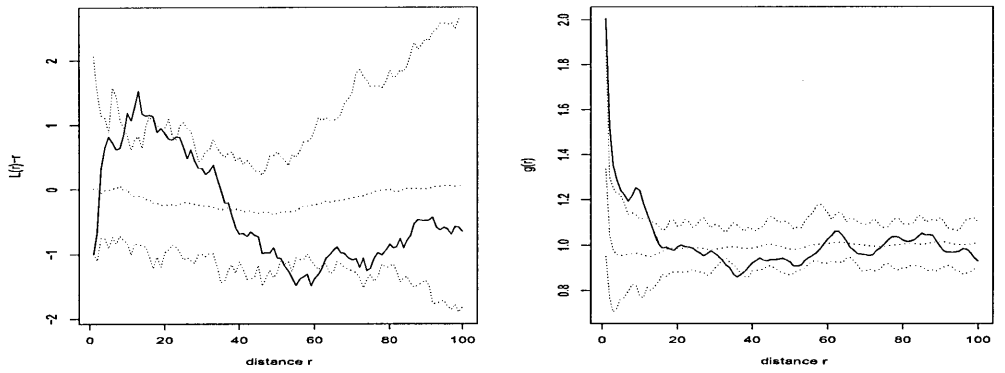

Figure B.74: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Leucopogon striatus*

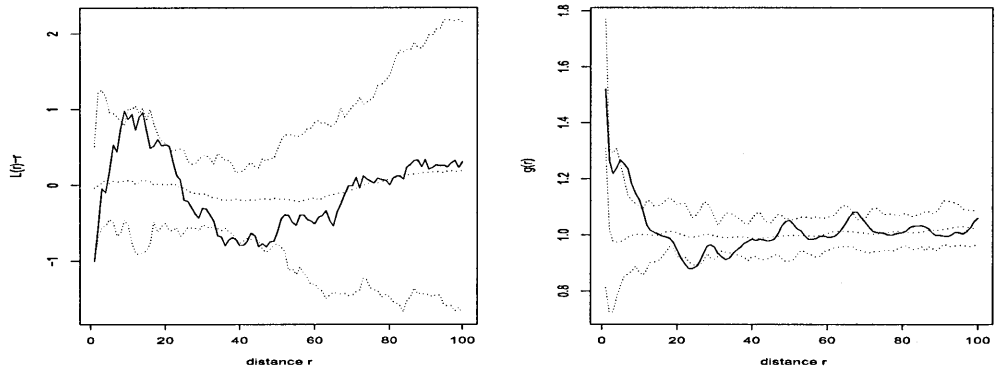

Figure B.75: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Lomadra sp.*

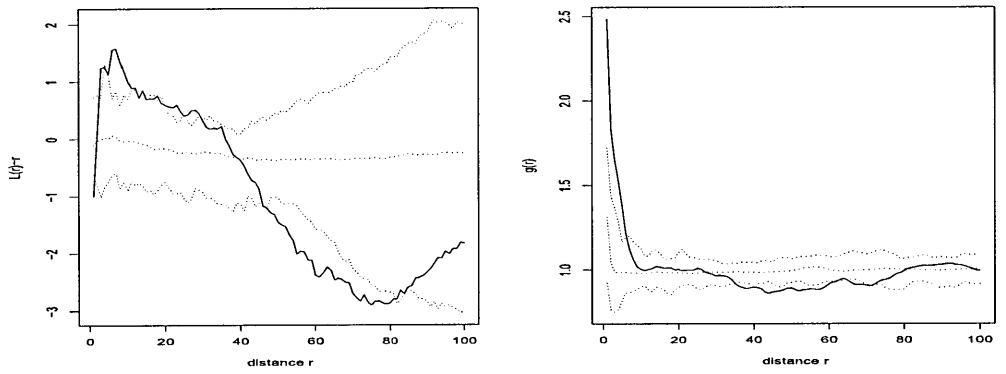

Figure B.76: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Lyginia barbata*

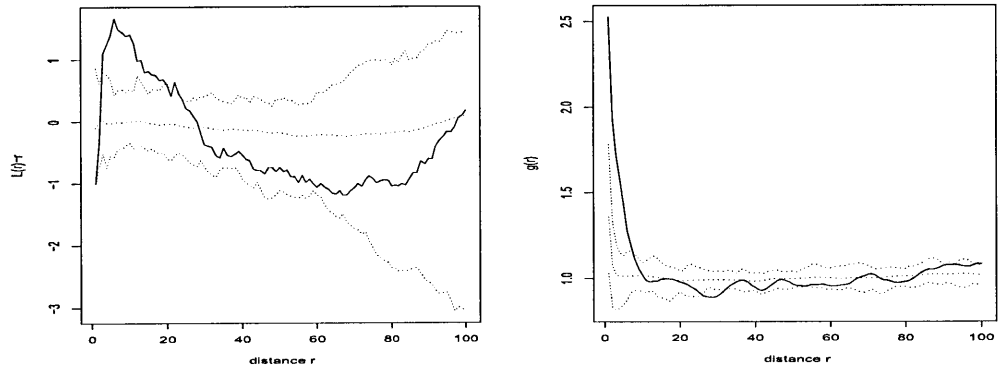

Figure B.77: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Melaleuca scabra*

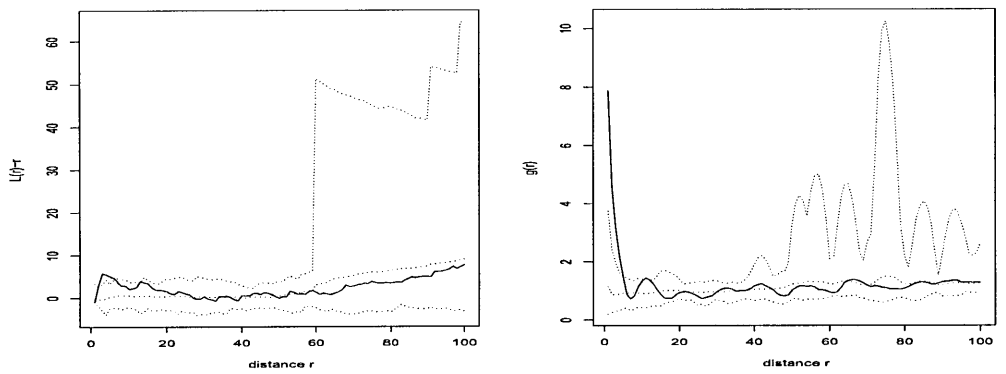

Figure B.78: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Patersonia occidentalis*

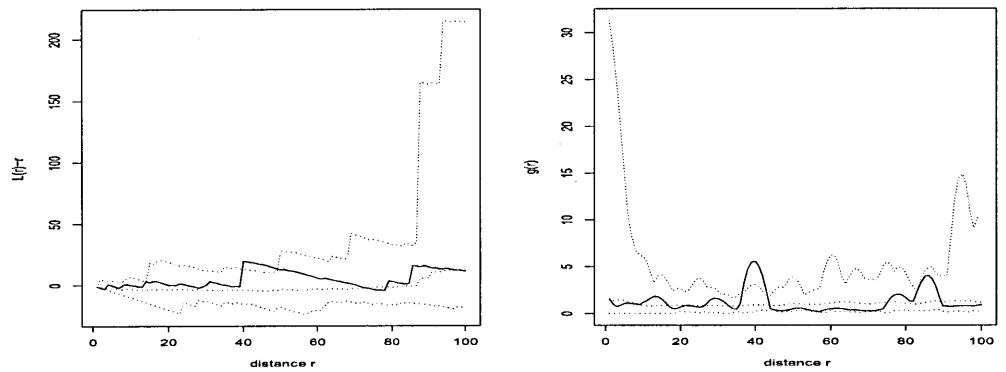

Figure B.79: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Platysace juncea*

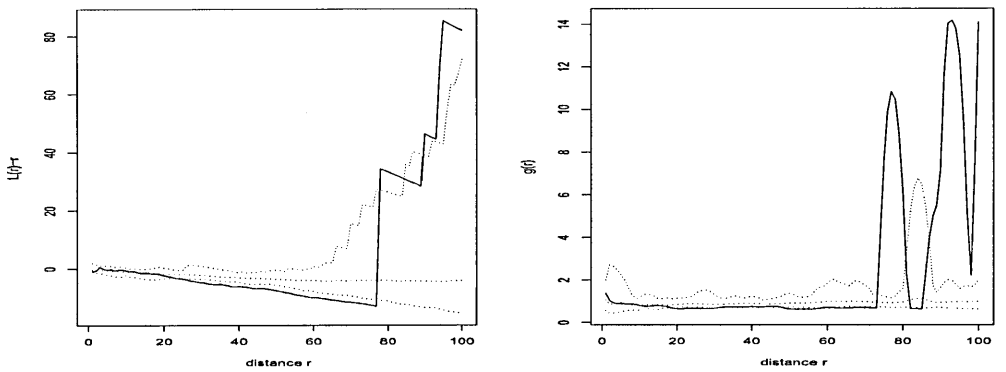

Figure B.80: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Chordifex sinuosus*

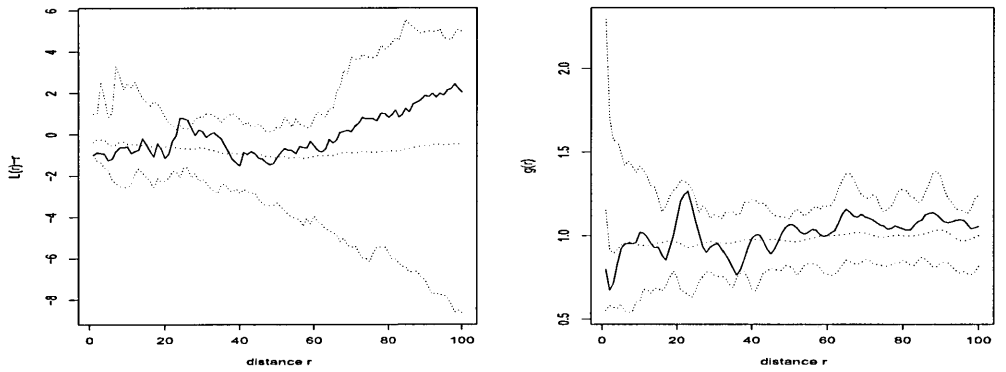

Figure B.81: Inhomogeneous L-function and pair-correlation function with simulation envelopes for *Scholtzia involucrata*

B.2.2 Results tests of complete spatial randomness

species	max-dist test	upper CUSUM	lower CUSUM
Adenanthos cygnorum	$p = 0.67$	$p = 0.09$	$p = 0.82$
Alexgeorgia nitens	$p = 0.02$	$p = 0.000$	$p = 1$
Andersonia heterophylla	$p = 0.000$	$p = 0.000$	$p = 1$
Astroloma xerophyllum	$p = 0.47$	$p = 0.000$	$p = 1$
Banksia attenuata	$p = 0.34$	$p = 0.20$	$p = 0.77$
Banksia menziesii	$p = 0.59$	$p = 0.22$	$p = 0.96$
Boronia ramosa	$p = 0.000$	$p = 0.000$	$p = 1$
Bossiaea eriocarpa	$p = 0.25$	$p = 0.01$	$p = 0.96$
Conospermum crassinervium	$p = 0.12$	$p = 0.000$	$p = 0.78$
Conostephium pendulum	$p = 0.000$	$p = 0.000$	$p = 1$
Conostylis canescens	$p = 0.04$	$p = 0.07$	$p = 0.62$
Conostylis juncea	$p = 0.23$	$p = 0.13$	$p = 0.91$
Dampiera linearis	$p = 0.48$	$p = 0.000$	$p = 0.98$
Dasypogon bromeliifolius	$p = 0.000$	$p = 0.000$	$p = 1$
Eremaea astrocarpa	$p = 0.25$	$p = 0.02$	$p = 0.32$
Eriostemon spicatus	$p = 0.45$	$p = 0.18$	$p = 0.99$
Hensmania turbinata	$p = 0.91$	$p = 0.000$	$p = 1$

Table B.3: Comparison of results from the max-dist test and the CUSUM approach for all species considered in this thesis, part I

species	max-dist test	upper CUSUM	lower CUSUM
Hibbertia crassifolia	$p = 0.64$	$p = 0.20$	$p = 0.87$
Hibbertia hypericoides	$p = 1$	$p = 0.54$	$p = 0.83$
Hibbertia subvaginata	$p = 0.400$	$p = 0.000$	$p = 1$
Hypocalymma xanthopetalum	$p = 0.25$	$p = 0.000$	$p = 0.95$
Isopogon linearis	$p = 0.61$	$p = 0.54$	$p = 0.89$
Jacksonia floribunda	$p = 0.000$	$p = 0.000$	$p = 1$
Chordifex sinuousus	$p = 0.000$	$p = 0.000$	$p = 1$
Lepidosperma angistatum	$p = 0.45$	$p = 0.000$	$p = 1$
Lepidosperma tenue	$p = 0.000$	$p = 0.000$	$p = 0.96$
Leucopogon conostephioides	$p = 0.81$	$p = 0.17$	$p = 0.76$
Leucopogon striatus	$p = 0.000$	$p = 0.000$	$p = 1$
Lomandra sp.	$p = 0.17$	$p = 0.000$	$p = 0.99$
Lyginia barbata	$p = 0.000$	$p = 0.000$	$p = 1$
Melaleuca scabra	$p = 0.14$	$p = 0.000$	$p = 1$
Patersonia occidentalis	$p = 0.05$	$p = 0.05$	$p = 0.94$
Phlebocarya filifolia	$p = 0.000$	$p = 0.000$	$p = 0.000$
Platysace juncea	$p = 0.83$	$p = 0.01$	$p = 0.99$
Scholtzia involucrata	$p = 0.000$	$p = 0.000$	$p = 1$
Stylidium crossocephalum	$p = 0.38$	$p = 0.17$	$p = 0.36$

Table B.4: Comparison of results from the max-dist test and the CUSUM approach for all species considered in this thesis, part II

Appendix C

Convergence plots for interaction parameters

We here present the convergence plots for the simulated interaction parameters derived from the MCMC algorithm. Here, plots that do not show a clear structure or pattern across iterations are indicative of approximate convergence of the Markov chains. All plots here cover the range of parameter values rather uniformly so that we can assume convergence.

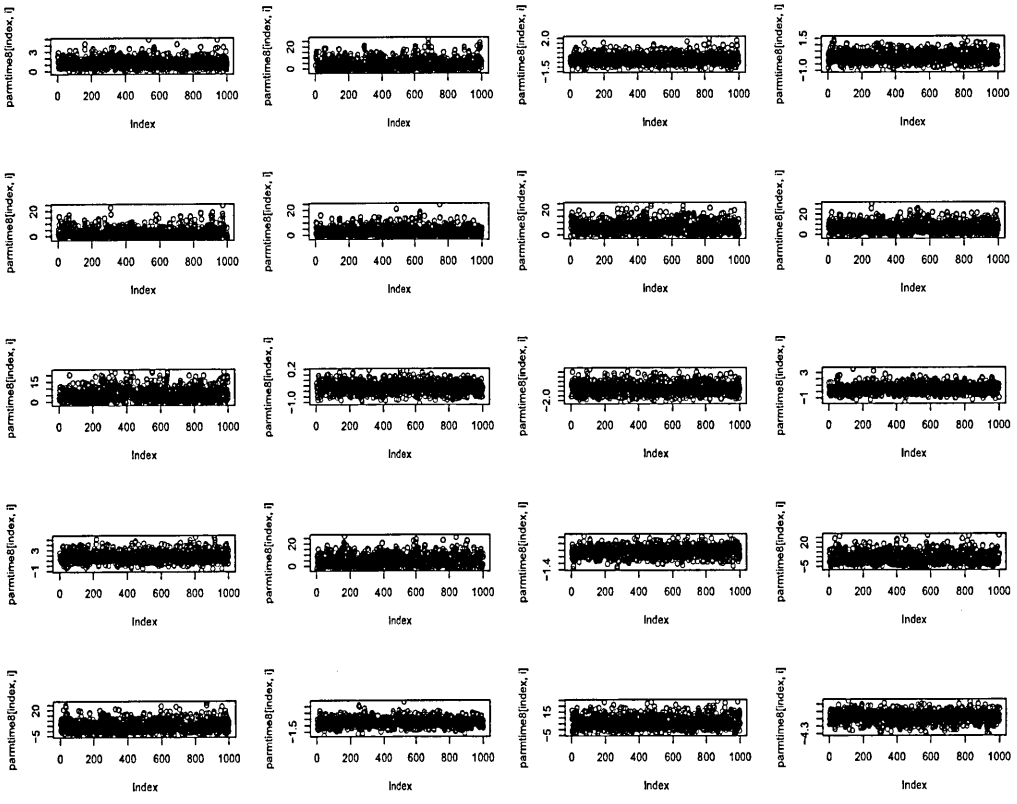

Figure C.1: Convergence plot for the interaction parameters for seeder species 1

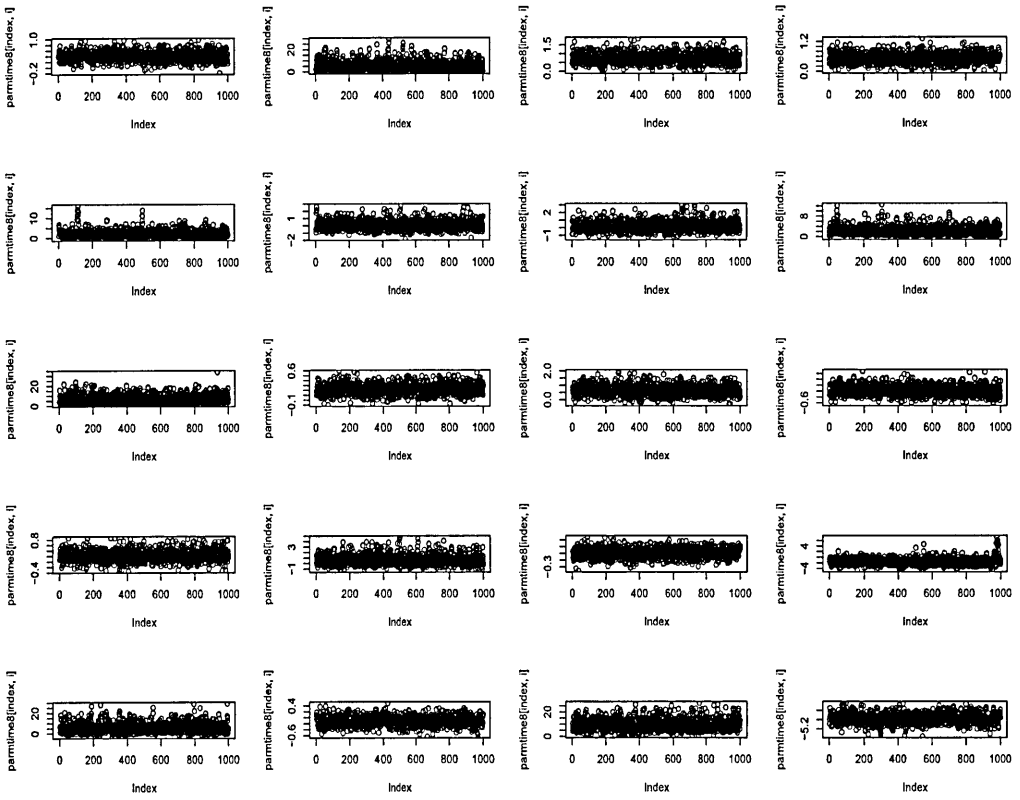

Figure C.2: Convergence plot for the interaction parameters for seeder species 2

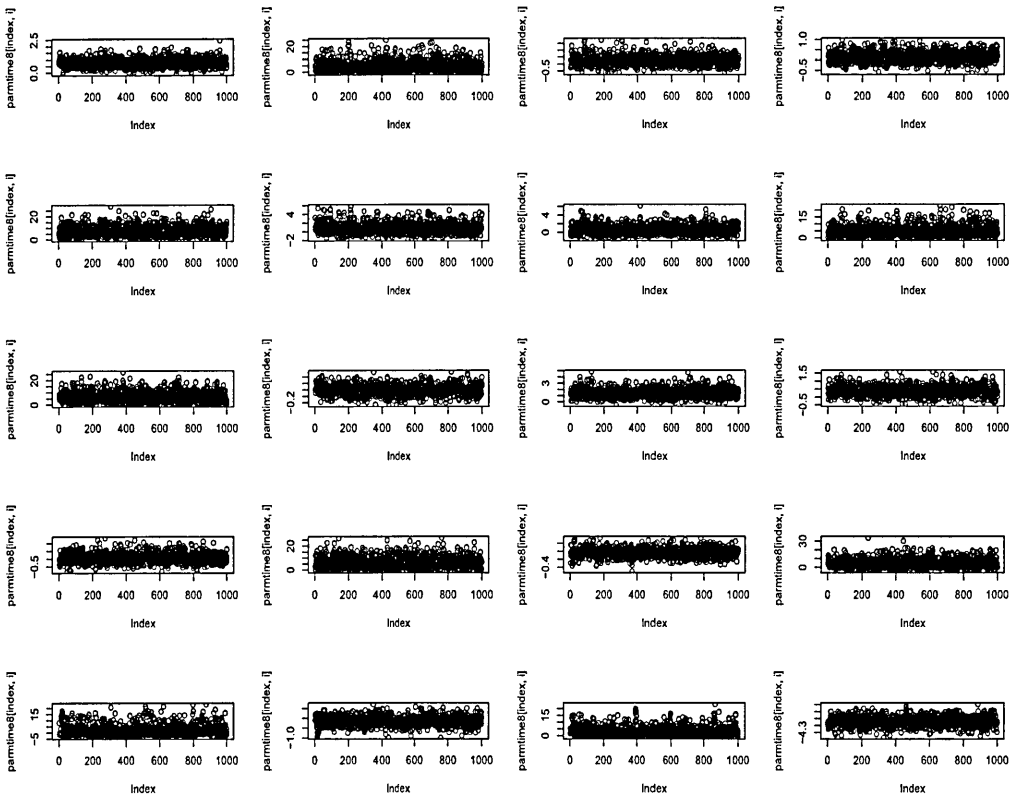

Figure C.3: Convergence plot for the interaction parameters for seeder species 3

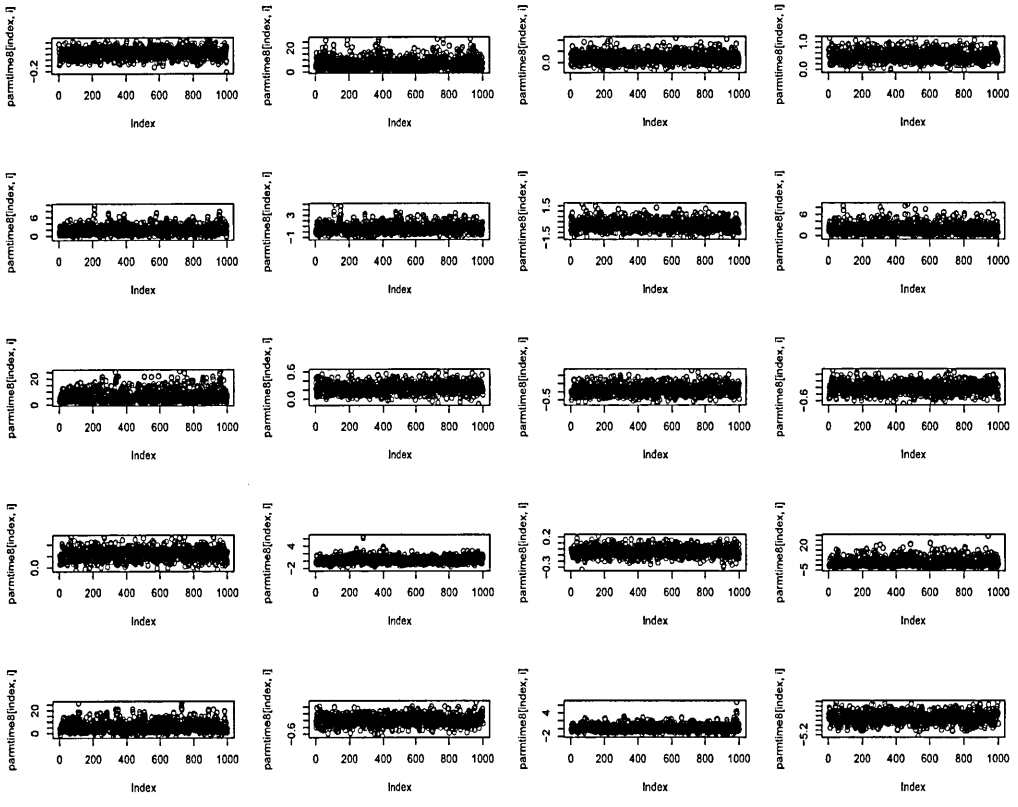

Figure C.4: Convergence plot for the interaction parameters for seeder species 4

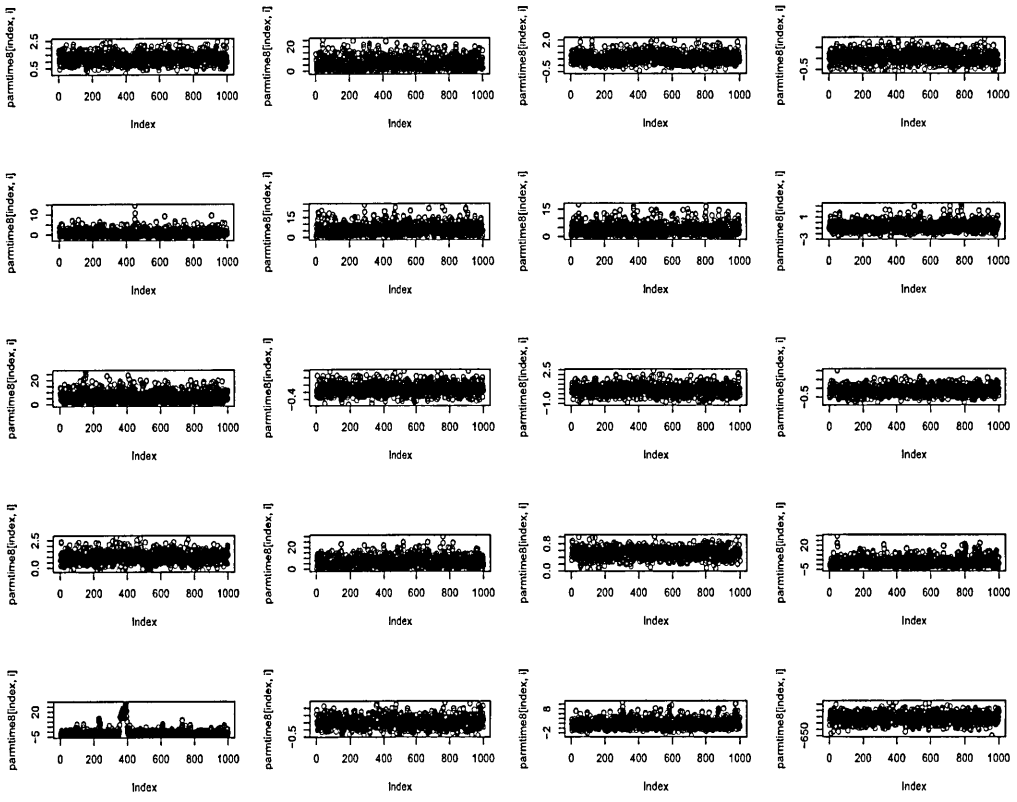

Figure C.5: Convergence plot for the interaction parameters for seeder species 5

Bibliography

Adler, R. (1981). *The Geometry of Random Fields.* Wiley, New York.

Amarasekare, P. (2003). Competitive coexistence in spatially structured environments: a synthesis. *Ecology letters 6*, 1109–1122.

Archibold, O. (1995). *Ecology of world vegetation.* Chapman & Hall, London.

Armstrong, P. (1991). Species patterning in the heath vegetation of the Northern Sandplain. Honours thesis, University of Western Australia.

Armstrong, P. (2005). Personal conversation.

Armsworth, P., B. Kendall, and F. Davis (2004). An introduction to biodiversity concepts for environmental economists. *Resource and Energy Economics 26*, 115–136.

Auld, T. and A. Denham (1999). The role of ants and mammals in dispersal and post-dispersal seed predation of the shrubs *Grevillea (Proteaceae). Plant Ecology 144*, 201–213.

Auld, T. and M. Tozer (1995). Patterns in emergence of *Acacia* and *Grevillea* seedlings after fire. *Proceedings of the Linnean Society of New South Wales 115*, 5–15.

Baddeley, A. and J. Møller (1989). Nearest-neighbour Markov point processes and random sets. *International Statistical Review 57*, 89–121.

Baddeley, A., J. Møller, and R. Waagepetersen (2000). Non- and semiparametric estimation of interaction in inhomogenous point patterns. *Statistica Neerlandica 54*, 329–350.

Baddeley, A. and B. W. Silverman (1984). A cautionary example for the use of second-order methods for analysing point patterns. *Biometrics 40*, 1089–1094.

Baddeley, A. and R. Turner (2000). Practical maximum pseudolikelihood for spatial point processes. *New Zealand Journal of Statistics 42*, 283–322.

Baddeley, A. J. and R. Turner (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software 12*, 1–42.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, Chichester New York.

Bascompte, J. and R. Sole (1998). *Modeling Spatiotemporal Dynamics in Ecology.* Springer-Verlag, New York.

Batista, J. and D. Maguire (1998). Modeling the spatial structure of tropical forests. *Forest Ecology and Management 110*, 293–314.

Bazzaz, F. (1996). *Plants in Changing Environments.* Cambridge University Press new York.

Beard, J. (1984). Biogeography of the kwongan. In J. Pate and J. Beard (Eds.), *Kwongan: plant life of the sandplains*, pp. 1–26. University of Western Australia Press, Nedlands.

Beard, J. (1990). *Plant Life of Western Australia.* Kangaroo Press Pty Ltd., Kenthurst, NSW.

Beard, J., A. Chaptman, and P. Gioia (2001). Species richness and endemism in the Western Australian flora. *Journal of Biogeography 27*, 1257–1268.

Bell, D. (1984). Aspects of response to fire in the northern sandplain heathlands. In J. Ford (Ed.), *Fire ecology and management in Western Australian ecosystems*, pp. 178–204. WAIT Environmental Studies Group Report No. 14, Bentley, WA.

Bell, D., A. Hopkins, and J. Pate (1984). Fire in the kwongan. In J. Pate and J. Beard (Eds.), *Kwongan: plant life of the sandplains*, pp. 1–26. University of Western Australia Press, Nedlands, WA.

Bell, G. (2001). Neutral macroecology. *Science 293*, 2413–2418.

Berman, M. and R. Turner (1992). Approximating point process likelihoods with GLIM. *Applied Statistics 41*, 31–38.

Berthelsen, K. and J. Møller (2002). A primer on perfect simulation for spatial point processes. *Bulletin of the Brazilian Mathematical Society 33*, 351–367.

Bertness, M. and R. Callaway (1994). Positive interactions in communities. *TREE 9*, 191–193.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician 24*, 179–195.

Besag, J. (1976). Some methods of statistical analysis for spatial data. *Biometrika 47*, 77–92.

Best, N., D. Spiegelhalter, A. Thomas, and C. Brayne (1996). Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society, Series A 159*, 323–342.

Bjørnstad, O., R. Ims, and X. Lambin (1999). Spatial population dynamics. *Trends in Ecological Evolution 14*, 427–431.

Bolker, B., S. Pacala, and C. Neuhauser (2003). Spatial dynamics in model plant communities: What do we really know? *American Naturalist 153*, 575–602.

Bond, W. and J. Midgley (2001). Ecology of sprouting in woody plants: The persistence niche. *Trends in Ecology and Evolution 16*, 45–51.

Bond, W., J. Vlok, and M. Viviers (1984). Variation in seedling recruitment of Cape *Proteaceae* after fire. *Journal of Ecology 72*, 209–221.

Bown, J., E. Pachepsky, A. Eberst, U. Bausenwein, P. Millard, G. Squire, and J. Crawford (2005). Consequences of intraspecifc variation for the structure and function of ecological communities i: Model development and predicted patterns of diversity (submitted).

Bradstock, R., M. Bedward, J. Scott, and D. Keith (1995). Simulation of the effect of spatial and temporal variation in fire regimes on the population viability of a banksia species. *Conservation Biology 10*, 776–784.

Brewer, M. (2000). A model-based approach for variable bandwidth selection in kernel density estimation. *Statistics and Computing 10*, 299–310.

Brix, A. (1999). Generalised gamma measures and shot-noise Cox processes. *Advances in Applied probability 31*, 929–953.

Brown, J. (1989). Regional variation in kwongan in the central wheatbelt of south-western Australia. *Australian Journal of Ecology 14*, 345–355.

Brown, J. and A. Hopkins (1983). The kwongan (sclerophyllous shrublands) of Tutanning Nature Reserve, Western Australia. *Australian Journal of Ecology 8*, 63–73.

Brown, N. (1993). Promotion of germination of fynbos seeds by plant-derived smoke. *New Phytologist 123*, 575–583.

Buckland, S., K. Newman, L. Thomas, and N. Koesters (2004). State-space models for the dynamics of wild animal populations. *Ecological Modelling 171*, 157–175.

Burslem, D., N. Garwood, and S. Thomas (2001). Tropical forest diversity – the plot thickens. *Science 291*, 606–607.

Burslem, D. and R. Law (2005). Analysis of tropical forest spatial pattern data: an overview for statisticians, unpublished document.

Cairney, J. and A. Ashford (2002). Biology of mycorrhizal associations of epacrids (ericaceae). *New Phythologist 154*, 305–326.

Cale, W., G. Henebry, and J. Yeakley (1989). Inferring process from pattern in natural communities – Can we understand what we see? *BioScience 39*, 600–605.

Callaway, R. M. (1995). Positive interactions among plants. *The Botanical Review 61*, 306–349.

Cardinale, B., A. Ives, and P. Inchausti (2004). Effects of species diversity on the primary productivity of ecosystems: extending our spatial and temporal scales of inference. *Oikos 104*, 437–450.

Cardinale, B., M. Palmer, and S. Collins (2002). Species diversity enhances ecosystem functioning through interspecific facilitation. *Nature 415*, 426–429.

Cardinale, B. J., K. Nelson, and M. Palmer (2000). Linking species diversity to the functioning of ecosystems: on the importance of environmental context. *Oikos 91*, 175–183.

Case, T. and R. Sidell (1983). Pattern and change in the structure of model and natural communities. *Evolution 37*, 832–849.

Caswell, H. (1976). Community structure: A neutral model analysis. *Ecological Monographs 46*, 327–354.

Chave, J. (2004). Neutral theory and community ecology. *Ecology Letters 7*, 241–253.

Chesson, P. (1985). Coexistence of competitors in spatially and temporally varying environments: A look at the combined effects of different sorts of variability. *Theoretical Population Biology 28*, 263–287.

Chesson, P. (1986). Environmental variation and the coexistence of species. In J. Diamond and T. Case (Eds.), *Community Ecology*, pp. 1–26. Harper and Row, New York.

Chesson, P. (2000). Mechanisms of maintenance of species diversity. *Annual Review of Ecology and Systematics 31*, 343–366.

Coates, D. (2000). Defining conservation units in a rich and fragmented flora: implications for the management of genetic resources and evolutionary processes in south-west australian plants. *Australian Journal of Botany 48*, 329–339.

Coates, D. and K. Atkins (2001). Priority setting and the conservation of Western Australia's diverse and highly endemic flora. *Biological Conservation 97*, 251–263.

Condit, R., P. Ashton, P. Baker, S. Bunyavejchewin, S. Gunatilleke, N. Gunatilleke, S. Hubbell, R. Foster, A. Itoh, J. LaFrankie, H. Lee, E. Losos, N. Manokaran, R. Sukumar, and T. Yamakura (2000). Spatial patterns in the distribution of tropical tree species. *Science 288*, 1414–1418.

Condit, R., N. Pitman, E. L. Jr., J. Chave, J. Terborgh, R. B. Foster, P. Nunez, S. Aguilar, R. Valencia, G. Villa, H. C. Müller-Landau, E. Losos,

and S. P. Hubbell (2002). Beta diversity in tropical forest trees. *Science 295*, 666–669.

Connell, J. (1971). On the role of natural enemies in preventing competitive exclusion in some marine animals and in rainforest trees. In P. den Boer and G. Gradwell (Eds.), *Dynamics of Populations*, pp. 298–313. Centre for Agricultural Publishing and Documentation, Wageningen, The Netherlands.

Coomes, D., M. Rees, and L. Turnbull (1999). Identifying aggregation and association in fully mapped spatial data. *Ecology 80*, 554–565.

Cox, D. (1955). Some statistical models related with series of events. *Journal of the Royal Statistical Society Series B 17*, 883–904.

Cox, D. and V. Isham (1980). *Point processes.* Chapman & Hall, London.

Cox, D., V. Isham, and P. Northope (2000). Statistical modelling and analysis of spatial pattern. In U. Dieckmann, R. Law, and J. Metz (Eds.), *The geometry of ecological interactions: Simplifying spatial complexity*, pp. 65–88. Cambridge University Press.

Crawley, M. (1997a). Biodiversity. In M. Crawley (Ed.), *Plant Ecology*, pp. 325–358. Blackwell Publishing, Oxford.

Crawley, M. (1997b). The structure of plant communities. In M. Crawley (Ed.), *Plant Ecology*, pp. 475–531. Blackwell Publishing, Oxford.

Cressie, N. (1991). *Statistics for Spatial Data.* Wiley, New York.

D'Agostino, R. and M. Stephens (1986). *Goodness-of-Fit Techniques.* Marcel Dekker, New York.

Dale (1999). *Spatial Pattern Analysis in Plant Ecology.* Cambridge Studies in Ecology, Cambridge University Press, Cambridge.

Dale, M., P. Dixon, M. Fortin, P. Legendre, D. Myers, and M. Rosenberg (2002). Conceptional and mathematical relationships among methods for spatial analysis. *Ecography 25*, 558–577.

Dale, M. and D. MacIsaac (1989). New methods for the analysis of spatial pattern in vegetation. *Journal of Ecology 77*, 78–91.

Daley, D. and D. Vere-Jones (1988). *An Introduction to the theory of point patterns.* Springer-Verlag, New York.

de Mazancourt, C. (2001). Consequences of community drift. *Science 293*, 1772.

DeAngelis, D. and L. Gross (1992). *Individual based Models and Approaches in Ecology: Populations, Communities and Ecosystems.* Chapman & Hall, London.

Dieckmann, U., R. Law, and J. Metz (2000). *The Geometry of Ecological Interactions – Simplifying spatial complexity.* Cambridge University Press, Cambridge.

Diggle, P. (1983). *Statistical Analysis of Spatial Point Patterns.* Academic Press, London.

Diggle, P. (2003). *Statistical Analysis of Spatial Point Patterns, 2nd ed.* Hodder Arnold, London.

Diggle, P., P. R. Jr, and O. Christensen (2003). An introduction to model-based geostatistics. In J. Møller (Ed.), *Lecture Notes in Statistics 173*, pp. 43–86. Springer-Verlag, New York.

Dixon, K. (2005). Personal communication.

Dryden, I. and K. Mardia (1998). *Statistical Shape Analysis.* Wiley, Chichester.

Duivenvoorden, J., J. Svenning, and S. Wright (2002). Beta diversity in tropical forests. *Science 295*, 636–637.

Dungan, J., J. Perry, M. Dale, S. Citron-Pousty, M.-J. Fortin, A. Jakomulska, P. Legendre, M. Miriti, and M. Rosenberg (2002). A balanced view of scaling in spatial statistical analysis. *Ecography 25*, 626–640.

Dunning, J., D.J.Stewart, B. Danielson, B. Noon, T. Root, R. Lamberson, and E. Stevens (1995). Spatially explicit population models: current forms and future uses. *Ecological Applications 5*, 3–11.

Durrett, R. and S. Levin (1998). Spatial aspects of interspecific competition. *Theoretical Population Biology 53*, 30–43.

Durrett, R. and S. A. Levin (1994). The importance of being discrete (and spatial). *Theoretical Population Biology 46*, 363–394.

Elkington, J. (1991). Report on the vegetation at Cooljarloo W.A., for TI02 Corporation. Ekomin Pty. Ltd., South Perth. Technical report.

Elton, C. (1927). *Animal Ecology.* Sidgwick and Jackson, London.

Enright, N., B. Lamont, and B. Miller (2005). Anomalies in grasstree fire history reconstruction for south-western Australian vegetation. *Australian Ecology 30*, 668–673.

Enright, N., R. Marsula, B. Lamont, and C. Wissel (1998a). The ecological significance of canopy seed storage in fire-prone environments: A model for non-sprouting shrubs. *Journal of Ecology 86*, 946–959.

Enright, N., R. Marsula, B. Lamont, and C. Wissel (1998b). The ecological significance of canopy seed storage in fire-prone environments: A model for resprouting shrubs. *Journal of Ecology 86*, 960–973.

Enright, N., B. Miller, N. Johnson, B. Lamont, and G. Perry (2004). Soil seed banks in three contrasting species-rich shrublands in south-western Australia. *Proceedings 10th MEDECOS Conference, Rhodes Greece.*

Ermentrout, G. and L. Edelstein-Keshet (1993). Cellular automata approaches to biological modelling. *Journal of Theoretical Biology 160*, 97–133.

Everitt, B., S. Landau, and M. Leese (2001). *Cluster Analysis.* Arnold, London.

Faith, D. (1994). Phylogenetic pattern and the quantification of organismal biodiversity. *Phylosophical Transactions of the Royal Society of London B 46*, 363–394.

Gaston, K. and S. Crown (2005). Editorial: Neutrality and the niche. *Functional Ecology 19*, 1–6.

Gause, G. (1934). *The struggle for existance.* Baltimore, MD: Williams and Wilkins.

Gelman, A., X. L. Meng, and H. S. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica 6*, 733–807.

Georgii, H.-O. and O. Haggström (1996). Phase transition in continuum Potts models. *Communications in Mathematical Physics 181*, 507–528.

Geyer, C. and J. Møller (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics 21*, 359–373.

Gill, A. (1996). Priorities for research and monitoring in mallee and heath lands of conservation reserves in northwestern Victoria. In A. Heislers, P. Lynch, and B. Walters (Eds.), *Fire ecology in semi-arid lands*, pp. 11–27. National Parks Service, Victoria.

Good, P. (2000). *Permutation Tests.* Springer, New York.

Green, P. and B. Silverman (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman & Hall, London.

Greig-Smith, P. (1983). *Quantitative Plant Ecology.* Blackwell, Oxford.

Griffin, E., A. Hopkins, and R. Hnatiuk (1983). Regional variation in mediterranean-type shrublands near eneabba, south-western Australia. *Vegetation 52*, 103–127.

Grinnell, J. (1917). The niche relationships of the Californian thrasher. *Auk 34*, 427–433.

Haase, P. (1995). Spatial pattern analysis in ecology based on ripley's k-function. *Journal of Vegetation Science 6*, 575–582.

Hanski, I. and M. Gilpin (1997). *Metapopulation biology: ecology, genetics and evolution.* Academic Press, San Diego.

Harms, K., R. Condit, S. Hubbell, and R. Foster (2001). Habitat associations of trees and shrubs in a 50-ha neotropical forest plot. *Journal of Ecology 89*, 947–959.

Hassell, M., H. Comins, and R. May (1991). Spatial structure and scale in insect population dynamics. *Nature 353*, 255–258.

Hector, A. (1999). Plant diversity and productivity experiments in european grasslands. *Science 286*, 1123–1127.

Herben, T., H. During, and R. Law (2000). Spatio-temporal patterns in grassland communities. In U. Dieckmann, R. Law, and J. Metz (Eds.), *The Geometry of Ecological Interactions: Simplifying Spatial Complexity*, pp. 11–27. Cambridge University Press, Cambridge.

Higgins, S., S. Picket, and W. Bond (2000). Predicting extinction risks for plants: Environmental stochasticity can save declining populations. *Trends in Ecology and Evolution 15*, 516–520.

Hoef, J. V. and N. Cressie (1996). Spatial statistics: Analysis of field experiments. In S. M. Scheiner and J. Gurevitch (Eds.), *Design and Analysis of Ecological Experiments*, pp. 319–341. Chapman & Hall, London.

Högmander, H. and A. Särkkä (1999). Multitype spatial point patterns with hierarchical interactions. *Biometrics 55*, 1051–1058.

Hopkins, A., G. Keighery, and N. Marchant (1983). Species-rich uplands of south-western australia. *Proceedings of the Ecological Society of Australia 12*, 15–26.

Hopper, S. (1979). Biogeographical aspects of speciation in the southwest australian flora. *Annual Review of Ecology and Systematics 55*, 1051–1058.

Hopper, S., M. Harvey, J. Chappill, A. Main, and B.York Main (1996). The western australian biota as gondwana heritage - a review. In S. Hopper, J. Chappill, M. Harvey, and A. George (Eds.), *Gondwana heritage: Past, present and future of the Western Australian biota*, pp. 11–27. Surrey Beatty and Sons, Chipping Norton, New South Wales.

Hubbell, S. (1979). Tree dispersion, abundance and diversity in a tropical dry forest. *Science 203*, 1299–1309.

Hubbell, S. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography.* Monographs in Population Biology 32, Princeton University Press.

Hughes, L. and M. Westoby (1992). Fate of seeds adapted for dispersal by ants in australian sclerophyll vegetation. *Ecology 73*, 1285–1299.

Huston, M., D. DeAngelis, and W. Post (1988). New computer model unify ecologcal theory. *BioScience 38*, 682–691.

Hutchinson, G. (1957). Concluding remarks. *Cold Spring Habor Symposium on Quantitative Biology 22*, 415–457.

Hutchinson, G. (1959). Homage to santa rosalia, or why are there so many kinds of animals? *American Naturalist 93*, 145–159.

Illian, J., E. Benson, J. Crawford, and H. Staines (2004). Multivariate methods for spatial point processes – a simulation study. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, and D. Stoyan (Eds.), *Spatial point process modelling and its applications*, pp. 125–130. Castelló de la Plana: Publicacions de la Universitat Jaume I.

Illian, J., E. Benson, J. Crawford, and H. Staines (2005). Principal component analysis for spatial point patterns. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, and D. Stoyan (Eds.), *Case studies in spatial point process modelling.* Springer, New York.

Illian, J., J. Møller, and R. Waagepetersen (2006). Spatial point process models for a complex plant community. *submitted*.

Jackson, R. and M. Caldwell (1993). The scale of nutrient heterogeneity around individual plants and its quantification with geostatistics. *Ecology 74*, 612–614.

Janzen, D. (1970). Herbivores and the number of tree species in tropical forests. *American Naturalist 104*, 501–528.

Jensen, J. and J. Møller (1991). Pseudolikelihood for exponential family models of spatial point processes. *Annals of Applied Probability 1*, 445–461.

Jolliffe, I. (2002). *Principal component analysis.* Springer.

Judson, O. (1994). The rise of the individual-based model in ecology. *TREE 9*, 9–14.

Kallenberg, O. (1984). An informal guide to the theory of conditioning in point processes. *International Statistical Reviews 52*, 151–164.

Keeley, J. (1991). Seed germination and life history syndromes in the california chaparral. *Botanical Review 57*, 81–116.

Kenny, B. (2000). Influence of multiple fire-related germination cues on three sydney grevillea (proteaceae) species. *Australian Ecology 25*, 664–669.

Kershaw, K. (1964). *Quantitative and Dynamic Ecology.* London: Edward Arnold.

Koen, C. (1991). Approximative confidence bounds for ripley's l-statistic for random points in a aquare. *Biometrical Journal 33*, 173–178.

Lamont, B. (1996). Conservation biology of banksias in southwestern australia. In S. Hopper, J. Chappill, M. Harvey, and A. George (Eds.), *Gondwana heritage: Past, present and future of the Western Australian biota*, pp. 11–27. Surrey Beatty and Sons, Chipping Norton, NSW.

Lamont, B., B. Collins, and R. Cowling (1985). Reproductive biology of the proteaceae in australia and south africa. *Proceedings of the Ecological Society of Australia 14*, 213–224.

Law, R., U. Dieckmann, and J. Metz (2000). Introduction. In U. Dieckmann, R. Law, and J. Metz (Eds.), *The geometry of ecological interactions: Simplifying spatial complexity*, pp. 1–6. Cambridge University Press, Cambridge.

Law, R., T. Herben, and U. Dieckmann (1997). Non-manipulative estimates of competition coefficients in grassland communities. *Ecology 85*, 505–517.

Law, R., D. Murrell, and U. Dieckmann (2003). Population growth in space and time: spatial logistic equations. *Ecology 84*, 252–262.

Law, R., D. Purves, D. Murrell, and U. Dieckmann (2001). Causes and effects of small scale spatial structure in plant populations. In J. Silvertown and J. Antonovics (Eds.), *Integrating Ecology and Evolution in a spatial context*, pp. 21–44. Blackwell Science, Oxford.

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology 74*, 1659–1673.

Lepš, J. (1990). Can underlying mechanisms be deduced from observed patterns? In F. Krahulec (Ed.), *Spatial processes in plant communities*, pp. 1–11. Academia Press, Prague.

Levin, D. (2000). *The origin, expansion, and demise of plant species.* Oxford University Press, New York.

Levin, S. (1992). The problem of pattern and scale in ecology. *Ecology 73*, 1943–1967.

Levins, R. and D. Culver (1971). Regional coexistance of species and competition between rare species. *Proceedings of the National Academy of Sciences, USA 68*, 12461–1248.

Liebhold, A. and J. Gurevitch (2002). Integrating the statistical analysis of spatial data in ecology. *Ecography 25*, 553–557.

Liebhold, A., R. Rossi, and W. Kemp (1993). Geostatistics and geographic information systems in applied insect ecology. *Annual review of Entomolgy 38*, 303–327.

Loreau, M. (2000). Biodiversity and ecosystem functioning: recent theoretical advances. *Oikos 91*, 3–17.

Loreau, M., S. Naeem, P. Inchausti, J. Bengtsson, J. Grime, A. Hector, D. Hooper, M. Huston, D. Raffaelli, B. Schmid, D. Tilman, and D. Wardle (2001). Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science 294*, 804–808.

Lotka, A. (1925). *Elements of physical biology.* Williams & Wilkins Co., Baltimore.

MacArthur, R. and E. Wilson (2001). *The Theory of Island Biogeography.* Princeton University Press.

Macdonald, I. (1985). The australian contribution to southern africa's invasive alien flora: An ecological analysis. *Proceedings of the Ecological Society of Australia 14*, 225–236.

Magurran, A. (1988). *Ecological Diversity and its Measurement.* University Press, Cambridge.

Mahdi, A. and R. Law (1998). On the spatial organisation of plant species in a limestone grassland community. *Journal of Ecology 75*, 459–476.

Marchant, N. (1973). Species diversity in the southwestern flora. *Journal of the Royal Society of Western Australia 56*, 23–30.

Mateu, J., J. Us'o, and F. Montes (1998). The spatial pattern of a forest ecosystem. *Ecological Modelling 108*, 163–174.

May, R. (1975). Patterns of species abundance and diversity. In M. Cody and J. Diamond (Eds.), *Ecology and Evolution of Communities*, pp. 81–120. Harvard University Press, Cambridge MA.

Meney, K., K. Dixon, and J. Pate (1997). Reproductive potential of obligate seeder and resprouter herbaceous perennial monocots (*Restionaceae, Anarthriacceae, Ecdeiocoleaceae*) from south-western Western Australia. *Australian Journal of Botany 45*, 771–782.

Meney, K., K. Dixon, and J. Pate (1999). Seed reproduction and germination ecology of restionaceae. In K. Meney and J. Pate (Eds.), *Australian rushes: Biology, identification and conservation of Restionaceae and allied families.* University of Western Australia Press, Nedlands, WA.

Møller, J., A. Syversveena, and R. Waagepetersen (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics 25*, 451–482.

Møller, J. and R. Waagepetersen (2003a). An introduction to simulation-based inference for spatial point processes. In J. Møller (Ed.), *Lecture Notes in Statistics 137*, pp. 143–198. Springer-Verlag, New York.

Møller, J. and R. Waagepetersen (2003b). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman & Hall/CRC, Boca Raton.

Morris, E. (2000). Germination response of seven east Australian *Grevillea* species (*Proteaceae*) to smoke, heat exposure and scarification. *Australian Journal of Botany 48*, 179–189.

Morris, E., A. Tieu, and K. Dixon (2000). Seed coat dormancy in two species of *Grevillea (Proteaceae). Annals of Botany 86*, 771–775.

Mouquet, N., J. Moore, and M. Loreau (2002). Plant species richness and community productivity: why the mechanism that promotes coexistence matters. *Ecology Letters 5*, 56–65.

Murrell, D., D. Purves, and R. Law (2001). Uniting pattern and process in plant ecology. *Trends in Ecology and Evolution 16*, 529–530.

Naeem, S. and L. Shibin (1997). Biodiversity enhances ecosystem reliability. *Nature 390*, 507–509.

Neuhauser, C. (2001). Mathematical challenges in spatial ecology. *Notes of the American Mathematical Society 11*, 1304–1314.

Noss, R. (1990). Indicators for monitoring biodiversity: A hierarchial approach. *Conservation Biology 4 4*, 355–364.

Ogata, Y. and M. Tanemura (1985). Estimation of interaction potentials of marked point-patterns through the maximum-likelihood method. *Biometrics 390*, 315–338.

Pacala, S., C. Canham, J. Saponara, J. Silander, R. Kobe, and E. Ribbens (1996). Forest models defined by field measurements: II. estimation, error analysis and dynamics. *Ecological Monographs 66*, 1–43.

Pachepsky, E., J. Bown, A. Eberst, U. Bausenwein, P. Millard, G. Squire, and J. Crawford (2005). Consequences of intraspecifc variation for the structure and function of ecological communities ii: Linking diversity and function (submitted).

Pachepsky, E., J. Crawford, J. Bown, and G. Squire (2001). Towards a general theory of biodiversity. *Nature 410*, 923–926.

Palmer, M. (1994). Variation in species richness: towards a unification of hypotheses. *Folia Geobot Phytotaxon 29*, 511–530.

Pate, J. and K. Dixon (1996). Convergence and divergence in the southwestern Australian flora in adaptations of roots to limited availability of water

and nutrients, fire and heat stress. In S. Hopper, J. Chappill, M. Harvey, and A. George (Eds.), *Gondwana heritage: Past, present and future of the Western Australian biota*, pp. 11–27. Surrey Beatty and Sons, Chipping Norton, NSW.

Perry, G., A. Liebhold, M. Rosenberg, J. Dungan, M. Miriti, A. Jakomulska, and S. Citron-Pousty (2002). Illustration and guidelines for selecting statistical methods for quantifying spatial patterns in ecological data. *Ecography 25*, 578–600.

Petchey, O., A. Hector, and K. Gaston (2004). How do different measures of functional diversity perform? *Ecology 85*, 847–857.

Petchey, O., P. McPhearson, T. Casey, and P. Morin (1999). Environmental warming alters food-web structure and ecosystem functioning. *Nature 402*, 69–72.

Phillips, O., P. Vargas, and A. Monteeagudo (2003). Habitat association among Amazonian tree species: a lanscape-scale approach. *Journal of Ecology 91*, 757–775.

Purdie, R. (1977). Early stages of regeneration after burning in dry sclerophyll vegetation. II Regeneration by seed germination. *Australian Journal of Botany 25*, 35–46.

Purves, D. and R. Law (2002). Fine-scale spatial structure in a grassland community: quantifying the plant's-eye view. *Journal of Ecology 90*, 121–129.

Purves, D. and R. Law (2003). Heteromyopia and the spatial coexistence of similar competitors. *Ecology letters 6*, 48–59.

Purves, D. and S. Pacala (2005). Ecological drift in niche-strucured communities: neutral pattern does not imply neutral process. In D. Burslem, M. Pinard, and S. Hartley (Eds.), *Biotic Interactions in the Tropics (in press)*. Cambridge University Press, Cambridge.

Ramsay, J. and B. Silverman (1997). *Functional data analysis.* Springer, New York.

Ramsay, J. and B. Silverman (2002). *Applied functional data analysis.* Springer, New York.

Read, D. J. (1996). The structure and function of the *ericoid* mycorrhizal root. *Annals of Botany 77*, 365–374.

Regan, H., R. Lupia, and M. Burgmann (2001). The currency and tempo of extinction. *American Naturalist 157*, 1–10.

Rice, J. and C. Wu (2001). Nonparametric mixed effect models for unequally sampled noisy curves. *Biometrics 57*, 253–259.

Richardson, D., R. Cowling, B. Lamont, and H. van Hensbergen (1995). Coexistence of Banksia species in southwestern Australia: the role of regional and local processes. *Journal of Vegetation Science 6*, 329–342.

Ripley, B. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability 13*, 255–266.

Ripley, B. (1988). *Statistical Inference for Spatial Point Processes.* Cambridge University Press, Cambridge.

Robert, C. P. and G. Casella (1999). *Monte Carlo Statistical Methods.* Springer-Verlag, New York.

Rossi, R., D. Mulla, A. Journel, and E. Franz (1992). Geostatistical tools for modelling and interpreting ecological spatial dependence. *Ecological Monographs 62*, 277–314.

Rummel, J. and J. Roughgarden (1983). Some differences between invasion-structured and coevolution-structured competitive communities: a preliminary analysis. *Oikos 41*, 477–486.

Sachs, L. (1984). *Applied Statistics: A Handbook of Techniques.* Springer Verlag, New York.

Schladitz, K. and A. J. Baddeley (2000). A third-order point process characteristic. *Scandinavian Journal of Statistics 27*, 657–671.

Schlather, M. (2001). On the second-order characteristics of marked point patterns. *Bernoulli 7*, 99–117.

Schneider, M., R. Law, and J. Illian (2006). Quantification of neighbourhood-dependent plant growth by Bayesian hierarchical modelling. *Journal of Ecology 94*, 310–321.

Silvertown, J., S. Holtier, J. Johnson, and P. Dale (1992). Cellular automaton models of interspecific competition for space – the effect of pattern on process. *Journal of Ecology 80*, 527–534.

Stoll, P. and J. Weiner (2000). A neighbourhood view of interactions among individual plants. In U. Dieckmann, R. Law, and J. Metz (Eds.), *The Geometry of Ecological Interactions: Simplifying Spatial Complexity*, pp. 11–27. Cambridge University Press, Cambridge.

Stoyan, D., W. Kendall, and J. Mecke (1995). *Stochastic Geometry and its Applications* (2nd ed.). John Wiley & Sons, London.

Stoyan, D. and A. Penttinen (2000). Recent applications of point process methods in forestry statistics. *Statistical Science 1*, 61–78.

Stoyan, D. and H. Stoyan (1994). *Fractals, random Shapes and Point Fields.* John Wiley & Sons, London.

Strauss, D. (1975). A model for clustering. *Biometrika 63*, 467–475.

Tilman, D. (1994). Competition and biodiversity in spatially structured habitats. *Ecology 75*, 2–16.

Tilman, D. and J. Downing (1994). Biodiversity and stability in grasslands. *Ecology 367*, 363–365.

Tilman, D. and P. Kareira (1997). *Spatial Ecology: The Role of Space in Population Dynamics and Interspecifc Interactions.* Princeton University Press, New Jersey.

Tilman, D., C. Lehman, and K. Thomson (1997). Population dynamics in spatial habitats. In D. Tilman and P. Kareiva (Eds.), *Spatial Ecology: The Role of Space in Population Dynamics*, pp. 3–20. Princeton University Press, Princeton.

Tilman, D., P. Reich, J. Knops, D. Wedin, T. Mielke, and C. Lehman (2001). Diversity and productivity in a long-term grassland experiment. *Science 294*, 843–845.

Tilman, D., Wedin, and J. Knops (1996). Productivity and sustainability influenced by biodiversity in grassland ecostystems. *Nature 379*, 718–720.

United Nations Environment Programme (1992). *Convention on Biological Diversity, NA 92-7807.* New York.

Upton, G. and B. Fingleton (1988). *Spatial data analysis by example. Volume 1.* John Wiley and Sons, New York.

Uriate, M., R. Condit, C. Canham, and S. Hubbell (2004). A spatially explicit model of sapling growth in a tropical forest: does identity of neighbours matter? *Journal of Ecology 92*, 348–360.

Uriate, M., S. Hubbell, R. John, R. Condit, and C. Canham (2006). Neighborhood effects on sapling growth and survival in a neotropical forest. In D. Burslem, M. Pinard, and S. Hartley (Eds.), *Biotic Interactions in Tropical Forests.* Cambridge University Press, to appear.

van Lieshout, M. (2000). *Markov point processes and their applications.* Imperial College Press, London.

van Lieshout, M. and A. Baddeley (1999). Indices of dependence between types in multivariate point patterns. *Scandinavian Journal of Statistics 26*, 511–532.

Volterra, J. (1926). Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem R Accad Nazionale del Lincei (Ser. 6) 2*, 31–113.

Waagepetersen, R. (2005). Posterior propriety for Poisson processes. Manuscript, http://www.math.aau.dk/ rw/publications.html.

Walter, H. (1973). *Vegetation of the earth in relation to climate and eco-physiological conditions.* (transl. J. Wieser), Springer Verlag, New York.

Walter, H. (2002). *Walter's Vegetation of the Earth: The Ecological Systems of the Geobiosphere.* (transl. G. Lawlor and D. Lawlor), Springer Verlag, Berlin.

Ward, D., B. Lamont, and C. Burrows (2001). Grasstrees reveal contrasting fire regimes in eucalypt forest before and after European settlements of southwestern Australia. *Forest Ecology and Management 150*, 323–329.

Watkinson, A. (1997). Plant population dynamics. In M. Crawley (Ed.), *Plant Ecology*, pp. 359–400. Blackwell Publishing, Oxford.

Weiss, M. (1907). L'hypothèse du champ moléculaire a la propropriété ferromagnétique. *Journal de Physique 6*, 661–690.

Wetherill, G. and D. Brown (1991). *Statistical Process Control: Theory and Practice.* Chapman & Hall, London.

Whelan, R. (1997). Seed dispersal in relation to fire. In D. Murray (Ed.), *Seed Dispersal*, pp. 359–400. Academic Press, Sydney, NSW.

Whittaker, R. (1972). Evolution and measurement of species diversity. *Taxon 21*, 213–251.

Widom, B. and J. Rowlinson (1970). A new model for liquid-vapor phase transitions. *Journal of Chemical Physics 52*, 1670–1684.

Wiegand, T. and K. Moloney (2004). Rings, circles, and null-models for point pattern analysis in ecology. *Oikos 104*, 209–229.

Wiens, J. (1989). Spatial scaling in ecology. *Functional Ecology 3*, 385–397.

Wilson, E. O. (1988). *Biodiversity.* National Academy Press, Washington.

Winkler, E., M. Fischer, and B. Schmid (1999). Modelling the competitiveness of clonal plants by complementary analytical and simulation approaches. *OIKOS 85*, 217–233.

Wright, S. J. (2002). Plant diversity in tropical forests: a review of mechanisms of species coexistence. *Oecologica 130*, 1–14.

Wu, J. and S. Levin (1994). A spatial patch dynamic modelling approach to pattern and process in an annual grassland. *Ecological Monographs 64*, 447–464.